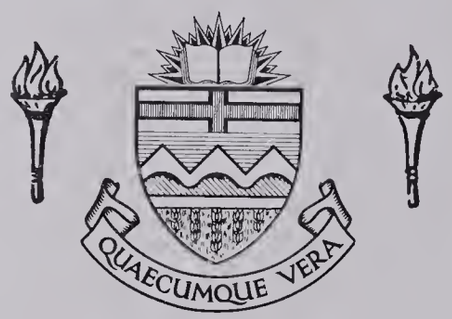
For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex libris
universitatis
albertaensis



THE UNIVERSITY OF ALBERTA

STUDY OF UDC AND OTHER INDEXING LANGUAGES THROUGH COMPUTER MANIPULATION OF MACHINE READABLE DATA BASES

by

© MARCEL ALBERT MERCIER

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

SPRING, 1972

Thesis
1972
101

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Study of UDC and Other Indexing Languages Through Computer Manipulation of Machine Readable Data Bases," submitted by Marcel Albert Mercier in partial fulfilment of the requirements for the degree of Master of Science.

Digitized by the Internet Archive
in 2019 with funding from
University of Alberta Libraries

https://archive.org/details/Mercier1972

ABSTRACT

Studies of information identification techniques used in document based information systems have been primarily concerned with the efficiency of the techniques when these are judged in terms of relevance/recall ratios. Such studies have also been concerned with methods of word control through a thesaurus, classification schedule, or subject heading authority. In general, these investigations have not been computer based. The recent advent of large machine readable files has lead to the possibility of new approaches in such investigations.

This thesis describes a study in which data bases, classification schedules, and a thesaurus, all in machine readable form, provide the tools for the investigation of the efficiency and suitability of a specialized thesaurus, traditional classification, and keyword indexing for the control of a document based information for water resource management. The MARC tapes, UDC schedules, UDC indexed data bases, a selected keyworded data base, and a water resource thesaurus are the processed elements. Through the use of such tools, methods are devised by which the suitability of information identification techniques can be determined for a particular system.

An on-line thesaurus controlled and classification related storage and retrieval system is suggested as the final goal. Problems of data tape format, file organization, coding and comprehensive systems design arise in the study.

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to my supervisor Mrs. Doreen Heaps, for her advice, criticism, and guidance throughout the duration of this research and to Dr. D. B. Scott, Chairman of the Department of Computing Science, for encouraging me in the pursuit of graduate study. In addition, I would like to express my gratitude to the following for their constant assistance and support: John Batteké, Fred Alber, Robert Freeman, Malcolm Rigby, Hans Wellisch, W. M. Schultz, Mrs. G. A. Cooke and her colleagues at the Boreal Institute.

This thesis was proofread by Donna Holland, and Mrs. Marlane Mercier, typed by Mrs. Marilyn Wahl and duplicated by Mrs. Isabel Bince. I also wish to express my appreciation to them.

This research was partially supported by grants and contracts from the Canada Department of Environment and the National Research Council of Canada.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 General Considerations

In the development of any document-based information system two outstanding problems have to be faced. These are, first, the drawing up of the design of the overall system and, secondly, the correct identification of the contents of the records of information. This thesis is concerned indirectly with the design of a large multi-disciplinary batch or on-line document based information service and directly with the choice of an appropriate computer assisted identification mechanism for the records in the system. Both problems, therefore, are of importance and the solution of the second will influence decisions taken in treating the first.

In both the choice of the identification techniques and in the implementation of the system it is proposed to employ new tools in the form of large machine readable data bases and to introduce new techniques using computer methods. These tools and techniques also influence the final design.

It is obvious that the overall design of an information system can be completed only after the difficulties of identification of information records have been solved; these difficulties are particularly acute for

an information system of the type to be discussed in this thesis. The type of data to be handled, the variety of users, and the incorporation of the new techniques and new tools all compound the difficulties. It is, therefore, considered necessary to outline briefly the problem of identification in a general way so that the particular issues and methods may be better understood.

1.2 Identification--Classification and Indexing

The problem of identification involves every aspect of the complex process of recognition. How do we recognize? How do we group like to like and separate similar from dissimilar? The need to recognize categories has existed from man's first efforts to control his society and to create order out of the chaos of information around him. Similarly, the defining of rules for such recognition has been the subject of learned discussion from the appearance of the first philosopher.

Information comes to man from many sources; sounds, scents, and tactile experiences are all in some way accounted for and categorized, discarded or used. Of these, written or spoken words are the primary means of organized communication. Formal information systems, ones in which recorded information is transmitted, operate through these written words. It is here taken for granted that the more efficiently these recorded words are categorized, the more efficient will be the transmission.

The history of man's attempts to order his written information records has been closely related to the discussions of order in the universe. Early classification schemes, that is the ordering of documents into classes ("classification"), were, naturally enough, indebted to philosophers such as Aristotle. These early attempts at classification primarily ordered material by examining characteristics and then by grouping like with like, with subsequent arrangement of the different groups in a hierachical structure in order of importance. As time went on, items were classified and then placed in a structure already given [1]. To carry out such a system of ordering and identifying written knowledge, at least two aids are necessary; a formal arrangement of hierarchies with an attached notation and a word authority list which leads into the hierarchies. Documents are assigned identification indicators (notation) according to their positions in the hierarchies.

Such an approach to document recognition and identification assumes that the hierarchies are more or less fixed. However, in the fullness of time, hierarchies disintegrate. At some point a newly encountered concept is recognized as unique. It does not belong in any of the established hierarchies. The unique characteristic is then more important in identification than the common characteristic.

In a retrieval system that retrieves documents through keyword access, a unique characteristic, such as distinctive keyword, will precisely identify one document and separate it from all others. This separation of documents on the basis of unique characteristics is generally achieved through "indexing", that is, by assigning descriptors or keywords to the document and is generally regarded as differing from "classification".

We may say, therefore, in a general sense, that classification of documents imposes order and insures control by trying to group them in like groups, but the indexing of documents imposes order and control by separating them on the basis of unlike characteristics. It is not surprising that the spectacular increase in indexing systems, whether uniterm or concept based, which came in the 1950's, should have appeared when established hierarchies governing set fields of knowledge were threatened and more precise documents were required to satisfy needs in new and specialized disciplines.

However, it must be recognized that keyworded, "indexed", documents fall into groups, that is to say, an implicit or hidden classification underlies the process of identifying documents by a string of unique words. The very recent increased emphasis on complete automated free text handling carries the same implicit definition [2]. It is assumed that a match will be made, that matching characteristics exist between free words in text. Equally, on

on the other hand, at the highest degree of specificity, uniqueness is implied in a classification scheme. For example, in some systems the uniqueness is supplied by the portion of the call number which indicates the individual author.

Neither the word authority list of the classification structure, the classification notation, nor keywords constitute the total words in a document. They are abstractions of the documents considered. It could be said that these words or class codes, if regularly assigned by some rules, either explicit or implicit, constitute meta languages. These presumably should have their own syntax; they are all "indexing" languages. This idea is not developed further at this point although this concept, discussed in a previous paper [3], underlies much of the approach taken in this thesis.

Despite at least two thousand years of discussion of classification, conflicts continue to exist between users of different classifications and between classifiers and indexers. It is true that each method offers unique advantages in the identification of information records. For example, indexing is more flexible than classification; but, classification is more structured than indexing. However, it has now become essential to resolve some of these conflicts and to denote attention to the definition and solution of basic problems. This step is necessary because certain fundamental questions have arisen in connection with

large scale contemporary information systems. The essential needs must be delineated and the conflicts resolved.

The most fundamental of these problems are the following:

1. Equating categories across different classifications. For example, all information on "Buddhism" would be placed in class 294.3 of the Universal Decimal Classification and in class BL1400-1485 of the Library of Congress Classification. How can equivalent classes be bridged?

2. Determining the most suitable existing classification for a given subject area, especially when a new collection is to be classified or indexed.

3. Combining techniques of indexing with techniques of classification. Readers should be aware of some related pioneering work carried out by the English Electric Thesaurofacet group [4].

4. Incorporating new techniques and new tools, especially those resulting from technological advances.

Solutions to these problems would markedly improve the transfer of information in large scale inter-disciplinary information systems. However, these problems have been difficult to solve previously because of the large manual effort that would have been required to perform tedious clerical tasks; there has been a lack of necessary tools to eliminate such manual effort.

This thesis attempts to help in the solution of the problems listed above. A generalized set of procedures is developed to help choose a classification, to combine indexing with classification, to relate classifications and to make use of new tools and techniques that have only recently become available.

These generalized procedures cannot be developed in a vacuum and, fortunately as this subject was under consideration, the opportunity arose to take part in the joint design of an information system for water resources planning. The author had been employed with the Water Planning and Operations Branch of the Federal Department of Environment where his primary concern was the dissemination of water resources information. In this branch it was becoming evident that a new information system had to be devised and that some logical method had to be found that could make use of various data bases in which information was processed under existing classification and indexing procedures. At the same time criteria had to be developed for the choice of a future classification or indexing language.

This thesis will therefore treat the problem in the context of a water resources information system with specific details being taken from the system at the Water Planning and Operations Branch. The following chapters define the Water Resources Information System and describe the implementation of the generalized procedures outlined above.

CHAPTER II

A WATER RESOURCES INFORMATION SYSTEM

Over the years, large amounts of document information and data have been generated by industry and government supported research and development projects in the field of water resources. These projects have arisen from public awareness of the need for competent water resource management in Canada. In order to effectively collect, manage, and distribute this information there must exist an efficiently designed Water Resource Information System.

2.1 Purpose

The purpose of such an information system would be to provide information required by all people involved with water resources development and management in Canada, in the most readily usable form, at the necessary time, and at the most convenient place. This information would include all data and intelligence needed for planning, coordinating, and administering water resource programs. The primary goal would be to make available to the Canadian scientific and engineering management community, research and other findings bearing on water resources.

2.2 Design Requirements

In effect, the system designed must ensure rapid access to all information related to the analysis and resolution of the problems of water resource management and planning in Canada. It must be concerned with the coordination of planning between federal and provincial levels, so often made more difficult than normal owing to inadequate facilities for the transfer of information.

In order to insure communication between federal and provincial levels, the system must be designed to operate on a cooperative basis. It should be based on the competent input from users in every province, each user employing the system as his own, while simultaneously enjoying the benefits of input from colleagues or input supplied by commercial agencies. This concept is depicted in Figure 1. It should be noted that the center for this information system would serve as a control unit to provide organization for the collection and dissemination of information gathered not only from abstracting agencies but from all users. Information assembled by one user would be available to all others.

2.3 Data to be Handled

The width and depth of coverage of scientific information of interest in the field of water resources will be outlined in this section. This is illustrated in Figure 2. The data that makes up this information can be distinguished into three basic types (See Figure 3).

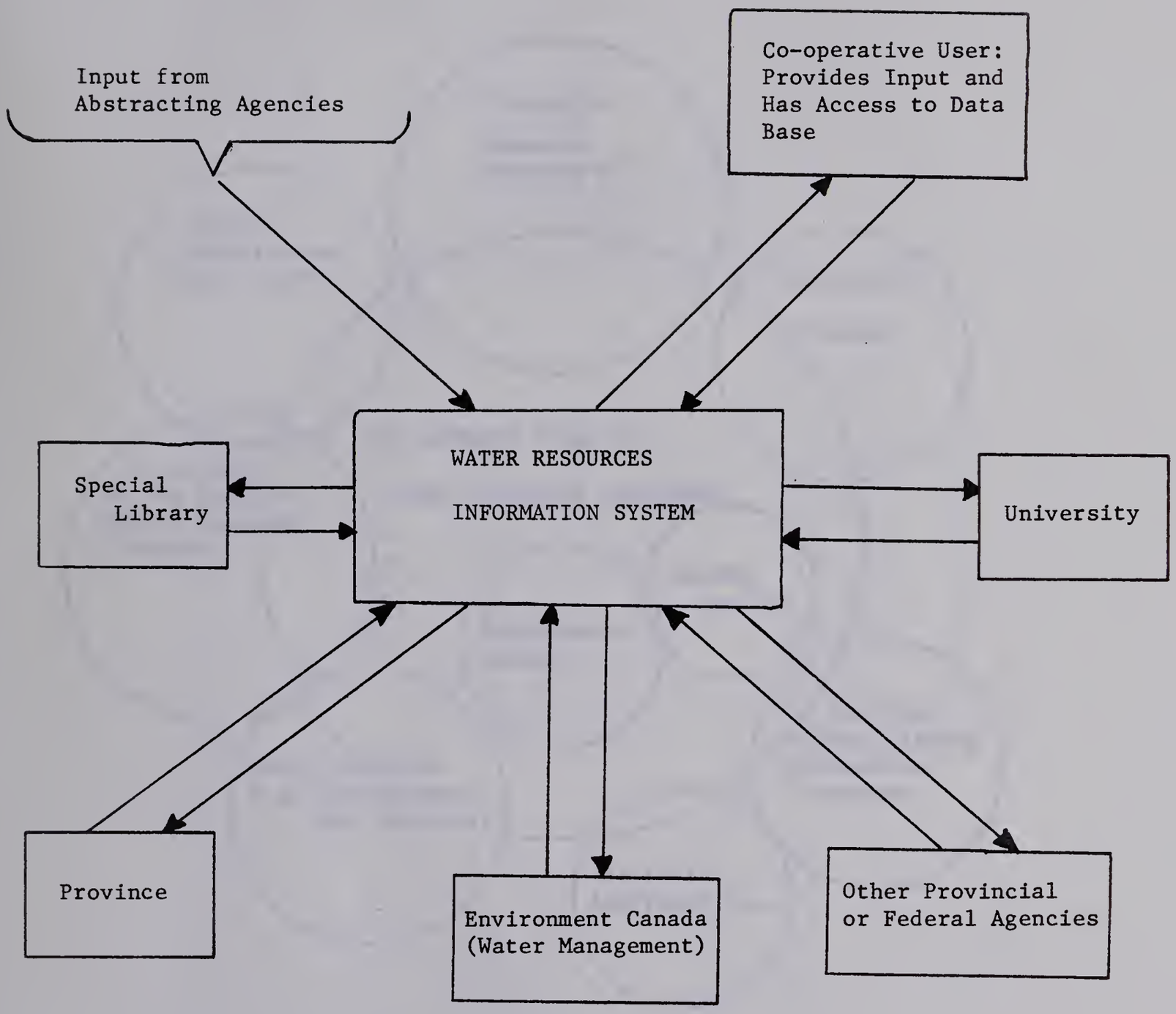


Fig. 1: Co-operative Users of a Water Resources Information System.

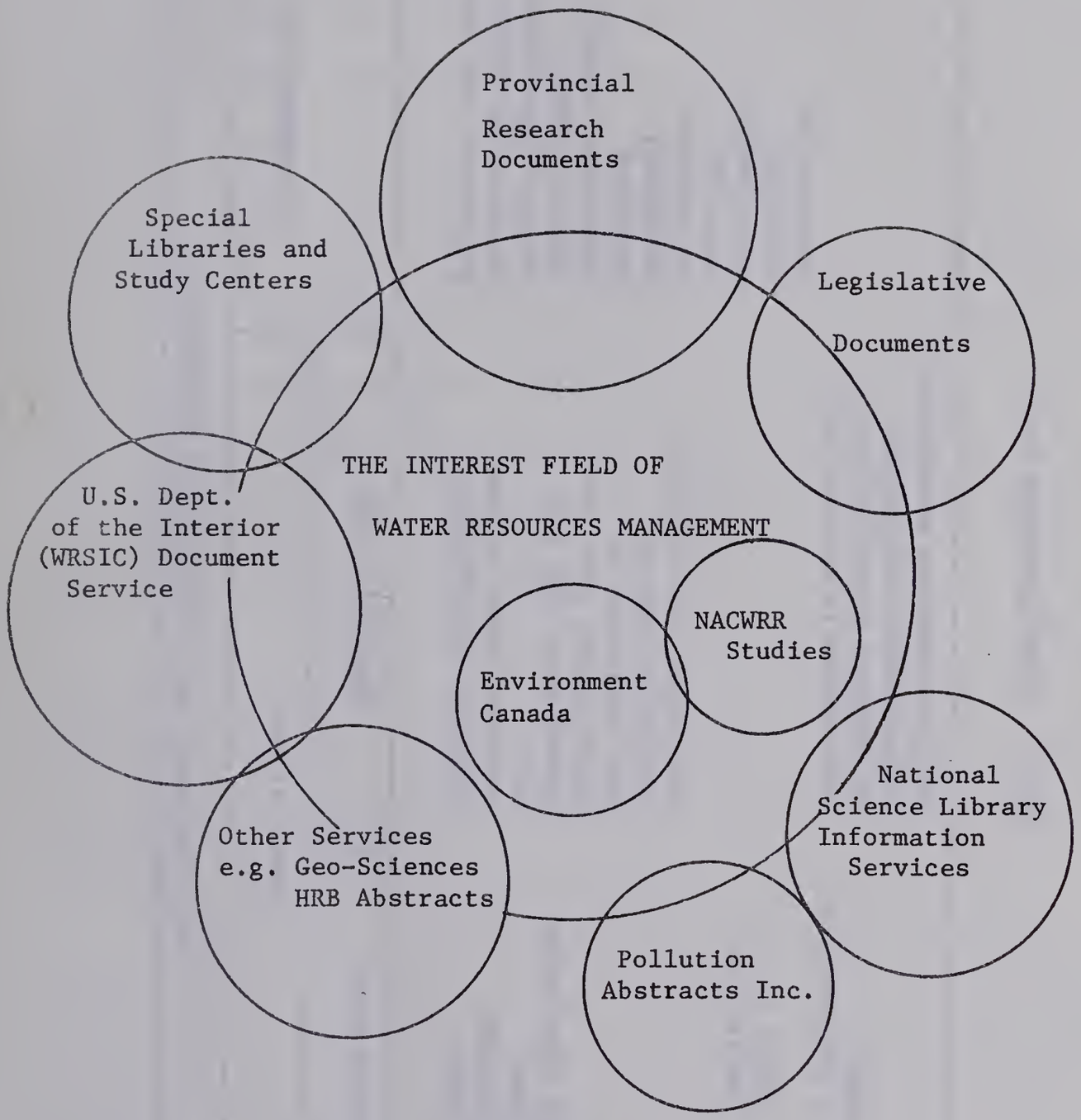


Fig. 2: Overlapping Interest Fields of Water Resources Management Information.

WATER RESOURCES DATA AND DOCUMENT INFORMATION		
Numerical Data		Documentation
High Quality Data	Incidental Data	
Statistical data: generally consistent and well defined, and containing time series For example: D.B.S. C.L.I. I.W.B. Water Quality Data I.W.B. Water Survey Application in mathematical models is possible	Produced in single-shot surveys, local studies, etc. Spread over time and geographical locations. Net consistent in definition and not available in time-series. Application in mathematical or economic models is only possible with significant modifications and assumptions	References that relate to Water Resources, covering literature on socio-economics analytical methods geography management legislation hydrology water quality pollution institutions

Fig. 3: Water Resources Data

These are:

1. "High quality" numerical data, such as that presently being used by the Dominion Bureau of Statistics (socio-economic), Canada Land Inventory (land use), Inland Waters Branch (water supply), and Water Planning and Operations Branch (water use). This consists mainly of high quality numerical data from which model building, econometric studies, and statistical analyses are conducted.

2. "Incidental" numerical data which consists mainly of lower quality, less consistent numerical information. This type of data is usually found in the form of tables or short, concise reports resulting from small or local studies and surveys.

3. Documentation. This consists of documentary or "word" or "nonnumerical" input, such as statutes and technical reports, administrative files, books, newspaper clippings, technical articles and so forth. This thesis is primarly concerned with the problems of this section.

In effect, a Water Resources Information System would deal with all types of relevant data and information, not only physical and techological, but also economic, sociological, political, and legal.

2.4 Users of a Water Resources Information System

The information system must meet the needs of a diversified group of users ranging from the highly sophisticated scientific and technical user to the general

non-technical. However, the primary group of users would comprise analysts (engineers, scientists, economists, sociologists, statisticians), at the Masters and Ph.D. level, who are involved with the management and development of water resources in Canada, both federally and provincially. They are the coordinators, demand forecasters, administrators and background policy makers.

2.5 Implementation

At first, the system designed would serve essentially the following branches of the Federal Department of Environment: the Water Planning and Operations Branch, the Inland Waters Branch, and the National Advisory Committee for Water Resources Research (NACWRR). It would be designed to supplement, not supplant, existing technical information services.

The system would offer a full range of services, some of which might be as follows:

1. Updated lists of research being done across Canada in the field of water resources would be provided. These lists would give details such as the work being done, by whom, where, starting date, and allotment of grants received.

2. Federal and provincial laws and statutes regulating water resources would also be available upon request.

3. Research reports, progress reports, or more generally, any information handled by the sections being served by this system would be made available to users.

If the system proved to be efficient and user studies warranted its usefulness, it could then be expanded to include material from and offer services to other interested government organizations, such as federal and provincial agencies, libraries, information centres and so forth. Eventually all agencies in Canada working in the area of water resources research and planning could take part in the information system without losing the power of making individual decisions.

2.6 Identification of Information--Classification and Control Vocabulary

In order to provide an effective Water Resources Information System an efficient identification scheme for the information to be handled must be developed. The development of such a scheme must be continuous and it must suit all persons and data related to water resource planning. To make the identification scheme as effective as possible classification will be combined with indexing and a controlled vocabulary will be used.

A controlled vocabulary is necessary and is particulary valuable in helping to reduce difficulties caused by the occurrence of synonyms and by ambiguity between terms. In fact, word investigation and control is the first step in the ordering of information.

It is evident that in building a library of documents, such as that required by a Water Resources Information System, the method used in identifying these documents will form the heart of the system. As must also be evident, from the previous discussion, if the data is arranged in a sensible and orderly manner it can also be retrieved in a sensible and orderly manner.

In a system that combines indexing with classification, the choice of an effective and suitable classification scheme should be the first step in the designing of the system. As stated, no method has previously been developed to help choose the classification scheme that would best suit a given data base. The following section will describe how such a method was developed for the system under consideration.

CHAPTER III

THE DEVELOPMENT OF A METHOD FOR CHOOSING A CLASSIFICATION SCHEME

The topic for this thesis, as previously stated, grew out of the need for a classification scheme for a non-numerical information data base for water resources. In the Water Planning and Operations Branch, formerly part of the Department of Energy, Mines, and Resources, now part of the Canada Department of Environment, a tentative classification scheme, which made use of a Water Resources Thesaurus, had been developed. Because of the inadequacy of the scheme, and the more complex requirements of a necessary larger system, it was decided to study other classification schemes. Those studies were to be on the following:

1. The Library of Congress (LC).
2. The Universal Decimal Classification (UDC).

The basic reason for studying these widely used classification schemes and the Water Resources Thesaurus is that at present, they are all being used in various installations for the classification and indexing of water resources information. Some of these are a) the Water Sector Library of the Canada Department of Environment, presently using the LC; b) the Water Resources Scientific Information Center of the United States Department of the

Interior, which makes use of the Water Resources Thesaurus; c) Water Planning for Israel, a governmental agency concerned with the design and planning of water resources both in Israel and abroad. In this agency, documents are indexed by the UDC and the Water Resources Thesaurus is also extensively used. It was considered of interest to include, if possible, the Dewey Decimal classification. This scheme is used in general purpose libraries.

An efficient method had to be developed that would help in deciding which of these classification schemes, if any, would be most appropriate for the classification of water resources information. A contract was awarded to develop such a method and to select the most appropriate scheme. Progress reports were to be submitted on a monthly basis. A large portion of the research reported in this thesis was carried out as part of the contract.

3.1 Planning the Method

In planning the method, five criteria were set down for the classification scheme to be chosen. There were:

1. The classification scheme must cover the entire area of water resources.

2. The scheme must be suitable for automated retrieval purposes.

3. It must be amenable to constant change, that is flexible.

4. It must be international.

5. The chosen classification scheme must be tied in with the Water Resources Thesaurus being used by the Branch. During the past several years, the thesaurus has been used extensively in the Water Planning and Operations Branch for documentation purposes. It has proved to be a reliable tool which covered the vocabulary in the field of water resources.

Programs to implement an on-line thesaurus based information storage and retrieval system which facilitates and makes use of this thesaurus-classification scheme linkage have been developed by F. Alber in conjunction with the work done in this thesis [5] [6]. A joint contract was awarded to support part of Alber's work.

No known method existed for testing a classification scheme to see if it would meet the above stated objectives. The Water Resources Thesaurus had not been linked in any way with the classifications to be studied. The algorithm developed in this thesis for testing classification gives not only a measure of the suitability for the classifications being studied but it also provides a tool for developing a concordance between a thesaurus and the classification being tested.

This algorithm makes use of the fact that concepts are represented by the vocabulary used in a field; the reasons for this approach should be clear from previous discussion. It was decided to match the vocabulary used in the field of water resources with the vocabulary used in the classifications to obtain a measure of the suitability of the classification

under test. Due to the very large number of words being dealt with, over one million, computer techniques had to be applied in implementing this procedure.

In order to apply computer techniques the classifications to be tested and the vocabulary used in the field of water resources would have to be available in machine readable form. This need was satisfied by obtaining large machine readable data bases that contained the necessary vocabularies. These data bases were obtained through an intensive search, which included extensive correspondence.

With the use of these newly acquired tools, a method which incorporated the idea of matching vocabulary against vocabulary, was developed for testing classification. The method resulted not only in obtaining a relative measure of the suitability of various classifications but also in a concordance between a thesaurus and a classification.

3.2 Project Plan

Although classification choice is the heart of the thesis, an entire information system was being designed, Therefore, the project plan, that was devised to keep control, had to allow for the time required in solving problems of overall system design. This project plan, as can be seen in Figure 4, consisted of four main sections to be worked at simultaneously with one another. These sections were:

1. System Definition.
2. UDC Testing.

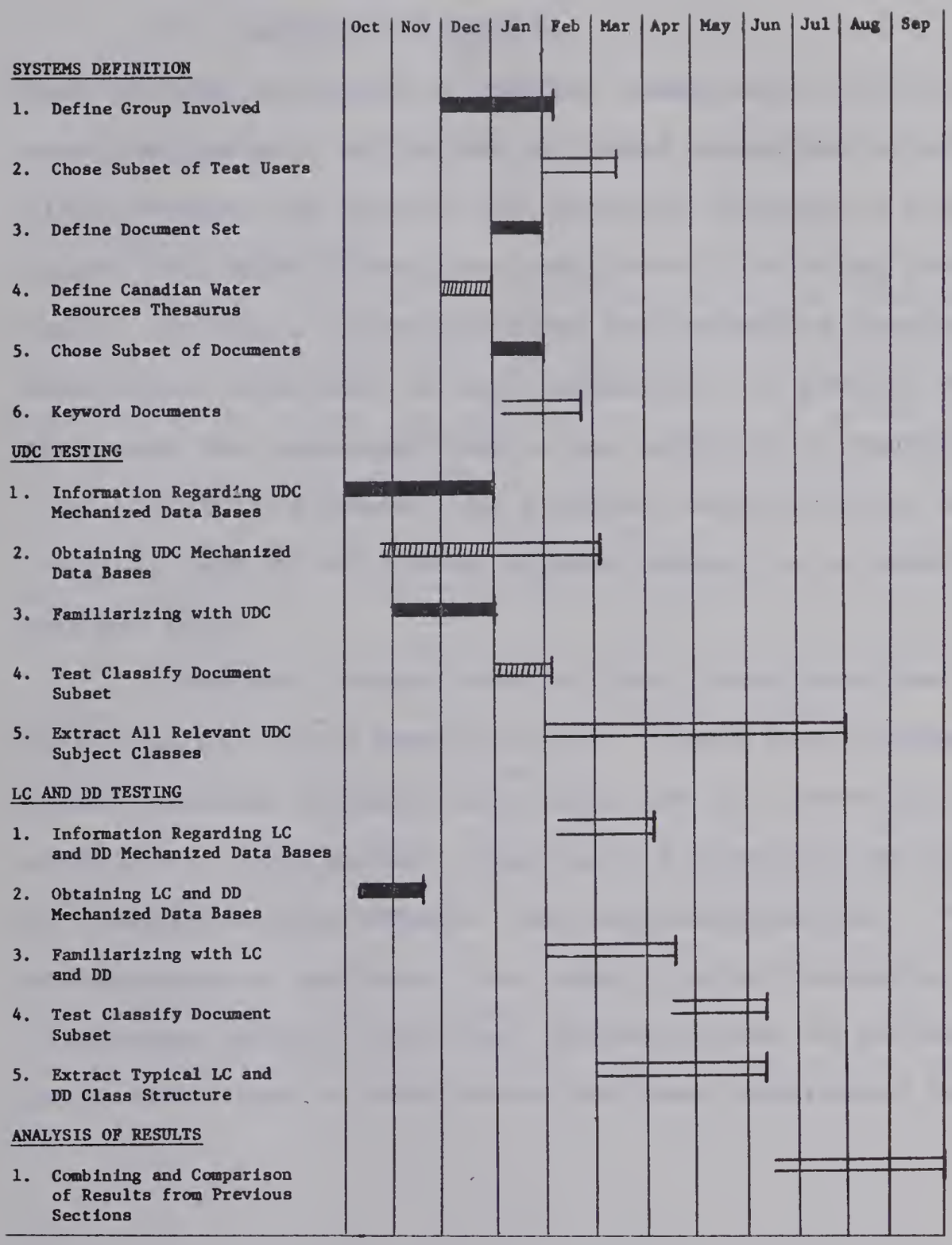


Fig. 4: Project Plan to January 1971

3. LC and DD Testing.

4. Analysis of Results.

Each section consisted of several tasks which will be described later. Definition of these tasks had to be kept fluid because the results and problems arising in one task helped determine techniques used later in solving other tasks. Difficulties encountered and impending results were often difficult, if not impossible, to predict because this work was concerned with a new approach to testing classification schemes. In a project such as this, an integral part of efficient systems design is an orderly project plan.

Numerous letters were written throughout the study in conjunction with every section. These were necessary to locate machine readable data bases and any other pertinent material or information, especially information on studies of classifications schemes (See Acknowledgements). This correspondence confirmed the results of an extensive literature search, which had indicated that no projects such as is described in this thesis had been undertaken before.

CHAPTER IV

PROCEDURE USED FOR CLASSIFICATION TESTING

4.1 General Comments

We have seen that no algorithm has been developed for classification testing. In the past, the choice of a classification scheme was often made on the basis of what scheme was most popular at the time, or of what scheme the developer of the system was most familar with. Often, poorly structured or disorganized schemes were developed without the slightest thought of studying already existing schemes. However, the lack of foresight shown is understandable for several reasons.

Over the years, classifications were generally found in large libraries and administered internally by librarians. As a result, very few users, professional or non-professional, are consciously aware of their existence. Even fewer people realize their advantages and the uses to which they can be put. It is no wonder that the non-librarian technologist or professional scientist or engineer, in Canada and elsewhere, who is often given the responsibility of collecting, organizing, and distributing information (that is providing an information system) does not know that the key to the system lies in the identification and consequent organization of the information.

As stated, this thesis is concerned with helping to solve the problem of organization of information by studying existing classification schemes through the manipulation of machine readable data bases. To be effective these data bases must include:

1. A large representative subset of the whole vocabulary used in a given subject field. It is through this vocabulary that a user of the system will express his ideas and communicate with the system.

2. The whole vocabulary used to retrieve information that has been classified. Note that this is not necessarily the vocabulary that is used by the classifier to classify information. It is the vocabulary that is provided by the classification for accessing the information that has been classified.

In this study, the Water Resources Thesaurus and the WPO word list provide the vacabulary that is used in the field of water resources. The UDC schedules provide the vocabulary that is used in accessing information from a UDC classified data base. The LC subject headings provide the vocabulary that is used in accessing information from an LC classified data base. These were obtained through an intensive search carried out by checking the literature, including conference announcements, news bulletins and similar material and through extensive correspondence resulting from all possible leads.

4.2 Methodology

The methodology developed for testing the classification scheme must serve two purposes. The first is that it must measure the appropriateness and adequacy of the scheme being tested for the particular information system. The second is that it should provide an effective tool for the uniting of the indexing scheme with the classification scheme. These two objectives are met by the use of the following method:

1. Define the standard vocabulary used in a given subject field.

2. Define the voacabulary used by users in retrieving information from a given classification scheme for that subject field. All words would have as tags the notational number of the subject class to which they lead.

3. Make the vocabularies from (1) and (2) available in machine readable form.

4. Compare one vocabulary with the other, keeping a record of all matches, including the number of the subject class.

5. Determine the percentage of words from (1) that were found to occur in the words from (2). This percentage will give a measure of the suitability of the classification scheme for a given subject field.

6. Print two lists. One list would be all matches ordered according to the words that were matched. The other

list would be all matches ordered according to the classification numbers.

7. Use the lists from (6) to check the relevancy of the matches.

The above definition of the method is very general. A detailed description of its application to the testing of the Library of Congress and the Universal Decimal Classification Schemes follows in the next two chapters.

As work proceeded, it was decided, at this time, to exclude the Dewey Decimal from the above described test for several reasons. The Dewey Decimal is general and suited to multi-purpose libraries rather than to special information systems. It is somewhat similar to the UDC and conditions were not favorable at this time to carry out a suitable test.

CHAPTER V

DATA BASES

The machine readable data bases acquired from various sources for use in this project are the UDC English Language Master File; the UDC updates to class 556, Hydrology; the United States Water Resources Scientific Information Center Thesaurus (WRSIC); the Geology Document File; Water Planning and Operations Branch (WPO) term list, WPO coded documents, and the MARC tapes (MAchine Readable Catalogue).

Several major difficulties were encountered with these data bases. Documentation was either very poor or non-existent. Errors were frequent and had to be accounted for in the programs written. Some of the errors were strictly format errors while others were due to programming errors that must have occurred when the tapes were created. Because of these problems, much time was spent in decoding the tapes. This is a tedious and time consuming task that frequently occurs when using data bases from outside the institution.

A description of the machine readable data bases obtained for this project will now be given.

5.1 UDC English Language Master Files

5.1.1 General Description

This file, on magnetic tape, includes the complete merged set of UDC schedules accumulated by the AIP/UDC

project from 1965 to 1967, a total of 110,759 records or approximately 1,000,000 words. Each record includes the UDC number, a code which represents the source of the record and the English language equivalent of the UDC number. Many records also contain cross references and scope notes, which serve to define and delimit a concept further.

5.1.2 Tape Format

All data is in character format (BCD).

All blocks and records are variable length.

Block format:

BL	Record 1	Record 2	...	Record n

BL is the block length. This can be a maximum of 1000 bytes including itself.

Record Format:

F1	F2	F3	F4	F5	V1 f1 f2 f3	...	V f1 f2 f3	$	Var1	Var2 ...	Var_n

F1: A 3 character fixed field. This field gives the number of characters in the record, including itself.

F2: A 19-character fixed field. This field contains the encoded UDC number, left justified.

F3: A 3 character fixed field. This field gives the code for the language of the data. "EN" in all AIP/UDC English Lanugage Files.

F4: A 4 character fixed field. This field gives the code for the UDC edition in which the record is published and the year of publication.

Example: AB61 = Abridged Edition, 1961.

F5: A 1 character fixed field. This field is unused and available for special uses.

Following these fixed fields is a Variable Field Map. This map consists of a variable number of 7 character fields composed of the following parts:

f1: 1 digit identifying the type of data (2=UDC heading, 3=scope note, 4=cross reference).

f2: 3 digits identifying the position of the first character of the variable field map, position 0 being a $ which separates the variable field map from the variable fields.

f3: 3 digits identifying the number of characters in the variable field.

The sources of input to the UDC English Language Master File and their codes are given in Table 1.

TABLE 1

SOURCES OF INPUT

ENAB61	Abridged Edition, UDC, 1961 complete
ENED65	Education Edition, 1965
ENFU--	Full Edition, unpublished ms
ENFU43	Full Edition, 1943 (partial)
ENFU54	Full Edition, 1954 (partial)
ENFU55	Full Edition, 1955 (partial)
ENFU58	Full Edition, 1958 (partial)
ENFU64	Full Edition, 1964 (partial)
ENME67	Medium Edition, 1967.

(EN is the language code meaning English)

These tapes were made available by Robert Freeman, who is a leading researcher of the UDC. The tapes were one result of the AIP/UDC project in which Mr. Freeman was a major coordinator [7].

5.2 UDC Updates

This file consists of the revisions that have been recently made to class 556, Hydrology. Mr. Geoffrey Lloyd, head of Classification Department, FID, provided a copy of these revisions. The updates were then keypunched and incorporated into this study. There were approximately 1500 words occurring in the updates.

5.3 Water Resources Thesaurus (WRT)

5.3.1 General Description

This thesaurus developed by the Water Resources Scientific Information Center, now part of the United States Department of the Interior, consists of 4039 main terms and their cross references interfiled with 1141 other terms (use references), that are intended to lead the user to the proper descriptors. A tape copy of this thesaurus was provided by the computer center at the University of Manitoba. The tape consists of an exact copy of the main body of the thesaurus; this was issued on November 1966.

5.3.2 Tape Format

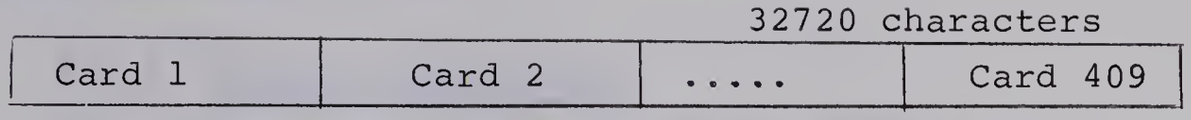
The tape format consists of card images blocked to lengths of 32720 bytes. Each of the main terms and use references are preceded by an asterisk. The main terms are followed by their cross references punched two per card in fixed fields starting at columns 4 and 44. Columns 1 and 40, of the cross reference cards, contain a code indicating whether the term is a broader term (BT), a narrower term (NT), a related term (RT), a used for term (UF), or a use reference (USE).

Example:

All data is in character format (EBCDIC).

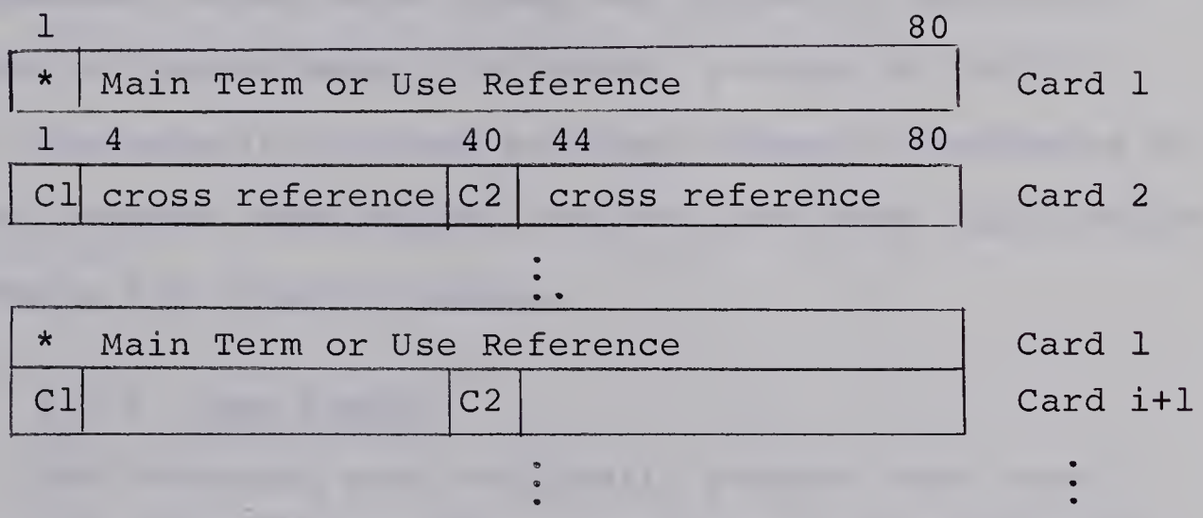
All blocks and records are fixed length.

Block Format:



Each block consisted of 409 cards

Record Format:



5.4 WPO Wordlist

A list of terms also incorporated in this study was supplied by the Water Planning and Operations Branch (WPO) of the Canada Department of Environment. This list is to complement the Water Resources Thesaurus. These new terms will, in effect, help Canadianize the thesaurus; the majority of terms are those that are relevant to Canadian policy formation and management of water resources.

These terms were punched onto cards; one term per card. There were 234 terms.

5.5 WPO Data Base

5.5.1 General Description

This file was comprised of approximately 500 documents that have been keyworded, abstracted and tentatively coded by the Water Planning and Operations Branch. Each of the documents, after being coded and placed in this file, consisted of approximately 200 words, a total of 100,000 words. The material covered by these documents consisted of research reports, monographs, statutes, and some applications and rewards for research grants.

5.5.2 Tape Format

The documents were originally punched onto cards, twenty-five cards per document. The file on the tape that was made available consisted of images of these cards. Blank coding forms and a copy of the coding manual were made available to help decipher the tape. These tools make the necessary tape format manipulation much easier.

5.6 Geology Document File

5.6.1 General Description

This is a tapecopy of a file of punched cards containing title and UDC numbers used in compiling the UNIDEK Index to Geoscience Abstracts (a part of the AIP/UDC project). There are approximately 7300 document entries; each entry consisting of approximately 10 words for the title and 50 characters for its encoded UDC number. This tape-copy was supplied by Robert Freeman.

5.6.2. Tape Format

All data is in character format (BCD).

All blocks and records are fixed length.

Block Format:

800 characters

Card 1	Card 2	....	Card 10

Each block consisted of 10 card images.

Record Format:

	61	78	80
DATA ELEMENT	C_1	C_2	C_3

DATA ELEMENT: The data element was either the title of the document or the UDC number for that document. When the data element was a UDC number it was sometimes enclosed in a code, *Z*DATA ELEMENT** or *Y*DATA ELEMENT**. This was a signal that the enclosed UDC numbers were to be retained in the original order, not permuted, for the purpose of a printed index.

C_1: This code gave an accession number.

C_2: This code gave the line sequence within the data element.

C_3: This code was the data element identifier. A '2' indicated that the data element was a title; a '4' indicated that the data element was a UDC number.

5.7 MARC Tapes

5.7.1 General Description

The MARC tapes consist of magnetic tapes distributed by the Library of Congress every two weeks on subscription. There are approximately 1000 titles per issue; they cover current monographs in all subject areas. Each entry is the image of its corresponding catalogue card. MARC tapes issued during 1970 were made available for use by the University of Alberta Library Systems Group.

5.7.2 Tape Format

A complete description of the MARC tapes, fully explaining the format used, can be found in the MARC manuals published by the Library of Congress [8]. Further information regarding format of MARC tapes for this project will be found in section 8.3.

CHAPTER VI

STUDY OF THE UNIVERSAL DECIMAL CLASSIFICATION (UDC)

The Universal Decimal Classification is a system of classifying information by analysis of idea content so that related concepts are grouped and subordinated [9] [10] [11] [12] [14]. It is an hierarchial, numerical classification scheme based on the Dewey Decimal principle that all knowledge is a whole and can thus be designated by 0., with infinite capability of subdivision as a decimal fraction.

It consists of:

1. A controlled and structured set of descriptors organized and displayed in a file with a variety of devices to facilitate comprehension of the structure and meanings.

2. A set of rules of formation, used to specify relationships among descriptors used to represent a given document. The scheme is regarded as universal in that an attempt is made to include in it every field of knowledge, not as a patchwork of isolated, self-sufficient specialist groupings, but as an integrated pattern of correlated subjects.

The UDC is based upon the major principle of general characteristics or facets. The term "facets" is used to represent a particular facet or aspect of a subject. If we consider a topic such as irrigation, we can place it in a

number of different contexts; for example it can be placed in agricultural hydraulics, a facet of hydraulic construction works, or it can be placed in rural engineering, a facet of agriculture, and so on. The major development of the UDC is continuously being made through the use of this principle. The theory of faceted classification will not be explained for this study. For anyone interested in this topic, consult The Elements of Library Classification, by S. R. Ranganathan [13] or Faceted Classification Schemes by Vickery [14].

As its name indicates, the Universal Decimal Classification is essentially a decimal classification, that is one in which each node of a tree of related descriptors may have up to ten nodes connected to it. However, this is not the case at one of the basic levels. The basic division of the stock of descriptors in UDC is into six facets, two form facets and four content facets, as illustrated in Figure 5.

Facets	Standard Notation
Form facets:	
1. language	=...
2. form of work	(0...)
Content facets:	
3. place	(n...) n=1 to 9
4. race	(=...)
5. time	'...'
6. general subject	absence of notational signal

FIG. 5: Form and Content Facets in UDC with Standard Notation

The ten primary divisions of the general subject facet, i.e. all knowledge, have been referred to as "main classes", while all other facets were known as "auxiliaries." These main classes can be expressed by single digits. (Since the "0" is common to all UDC numbers it is usually omitted.)

0. Generalities--Libraries, etc.
1. Philosophy, Psychology
2. Religion. Theology
3. Social Science--Law
4. Vacant at present
5. Natural Sciences
6. Applied Science. Technology
7. Arts. Architecture. Sport
8. Linguistics. Languages. Literature
9. Geography. Biography. History.

Each of these main classes is then further subdivided decimally to the required degree. Each subdivision, or subject class, is assigned a particular UDC number; the more specific the subject field, the longer the number.

Example:

6	Applied Science
62	Engineering and Technology
621	Mechanical and Electrical Engineering
621.1	Steam Power
621.11	Steam Power Plants

However, this is not always true because sometimes, in order to shorten notation, a single subject may be spread over several divisions.

All subject classes may also be combined through the use of various symbols. An example of these symbols are:

+ coordination--connects two distinct subjects discussed together

: subordination--separates two distinct main numbers to classify one work

/ connects distinct subjects having consecutive numbers

Analytical subdivisions may also be applied to subject classes through the use of the following symbols:

- or .0	Analytics
.00	Point of view

For convenience the initial decimal point of a class number is omitted; but, in most cases there is a decimal placed after every third digit. For example, 628.1 is the UDC number for Water Supply and Water Works. The initial 6 is classified under applied science and the leading decimal point has been deleted.

Every number, assigned to a subject, is included (by application) in a Parent Number, so that the degree of precision in classifying may be adjusted to the needs of the individual user. For example 628.1 is included in its parent number 628.

A book or document is classified according to the particular fields of interest that it deals with. It may be assigned one or more of the general subject numbers as well as any auxiliary numbers.

In forming a document index by using UDC, emphasis has been placed on keeping a definite order in stating the descriptors to be used to represent a document and on a definite order for filing such records. This means that descriptors from auxiliaries have not been able to serve as direct points of entry to the document index. The only apparent reason for this practice has been the necessity to limit severely the number of entries in a manual card file for reasons of cost and sheer physical size. In a manual file, one card is needed for each desired entry point to a document. An example of the full use of the UDC indexing for one document is given in Figure 6. Note that in a

manual file nine cards would be needed if access to this document was required for each of the entry points.

Document Title: Distribution and Seasonal Movements of Saginaw Bay Fishes	
UDC in statement form: 597:591.9:591.52(285:71:73)"1964/1966"(047+084.3)	
UDC in index entry with English language equivalents:	
(047)	Technical Reports
(084.3)	Maps
(285)	Lakes
(71)	Canada
(73)	United States
"1964/1966"	Events of 1964-1966
591.52	Animal Habitats and Migrations
591.9	Geographical Distribution of Animals
597	Ichthyology. Fish

Fig. 6: Example of UDC Indexing of a Document [15]

No theoretical reason makes any of the entries shown in Figure 6 auxiliary to any other and incapable of serving as a valid point of entry to a file. In a mechanical file all of these entries could be considered as valid entry points to an index number. That is, any document index number could be entered at any point once it had been stored in the data base.

6.1 Notation

UDC and other indexing languages which make use of a non-natural language notation are frequently criticized on the ground that such notations introduce unnecessary complications for the user. This is not a valid criticism

when considering the suitability of the notation in a mechanized system. The multiplicity of access points offered by a properly worked out notation is an asset. Moreover, many automated systems convert natural language descriptors to a non-natural language descriptor surrogate or code at the input stage. The latter is then used for all internal processing and retrieval purposes. For automatic retrieval systems, the advantage would appear to go to the indexing language with a notation that not only uniquely identifies the descriptors, but reveals relationships among them. This notation, to be most useful, should reveal relationships between descriptors used to index a particular document.

Regardless of how the notation is derived and assigned, it is necessary for the system designer and programmer to understand the significance of the notation and the method by which it should be handled in a mechanized retrieval system.

The UDC notation consists of strings of digits, interspersed for visual convenience with decimal points, at intervals of three digits, and other symbolic indicators. Various symbols occur which indicate the facet to which the descriptor belongs and to signal subordinate descriptors. In addition some symbols exist that can act as connectives to specify certain relationships between two UDC numbers when used to describe the complex content of a document.

6.2 Coding the Notation for Computer Manipulation

The UDC notation is based partially on the need for visual convenience, a meaningless function in a computer. In coding the notation for computer manipulation it is more efficient to store the notation in the computer in a form that preserves the concepts and relationships of the UDC, but uses a somewhat different set of symbols for internal processing reasons. Some of these reasons may be as follows:

1. Some of the UDC symbols use the same punctuation symbols. The order of sorting these characters, arbitrarily defined for each type of computer, causes some problems when they are interspersed among digits in UDC numbers, especially if the user decides to adhere to UDC filing rules.

2. Some of the UDC symbols have different meanings when used in different contexts. For example, the equal sign has one meaning by itself and another if preceded by a parenthesis.

3. Two of the UDC symbols (0. and .00) incorporate a meaningful use of what is otherwise a convenience symbol, the decimal point.

Any arbitrary set of symbols may be used for the efficient coding of the notation in a computer. The chosen set depends on the computer system being used and the criteria set down by its developers. A set of symbols being used for the storage of the UDC schedules on magnetic tape, resulting from the AIP/UDC research project, appear in Table 2.

TABLE 2

UDC INDEX NUMBERS ENCODED FOR STORAGE
AND RETRIEVAL BY COMPUTER

TYPE	NAME	NORMAL FORM	ENCODED FORM	EXAMPLE	ENCODED FORM
CONTENT FACET	GENERAL SUBJECT	n	nC	551.525	551Q525C
FORM FACET	LANGUAGE	=n	Fn	=30	F30
CONTENT FACET	PLACE	(mn)	HmnH	(265)	H265H
CONTENT FACET	RACE	(=n)	JnH	(=30)	J30H
SUBORDINATE CONTENT FACET	POINT OF VIEW	.00n	Ln	55.002	55L2C
SUBORDINATE CONTENT FACET	SPECIAL AUXILIARY	-n	Mn	62-451	62M451C
SUBORDINATE CONTENT FACET	SPECIAL AUXILIARY	.On	Nn	62.018.7	62N18Q7C
CONNECTIVE	SYNTHETIC CONNECTIVE	n'n	nPn	546.3'13	546Q3P13C
CONNECTIVE	INCLUSIVE CONNECTIVE	n/n	nBn	543/546	543B546
CONNECTIVE	RELATIVE CONNECTIVE	n:n	nDn	543:546	543D546
CONNECTIVE	GENERAL CONNECTIVE	n+n			
CONNECTIVE	SUBORDINATE CONNECTIVE	n(n)	nHnH	543(546)	543H546H

NOTE: n = SET OF DIGITS, ANY OF WHICH MAY BE ANY OF THE DIGITS 0....9

m = a DIGIT FROM 1....9

The rules required for transformation in this table should be self evident and should not require formal statement. As previously mentioned, due to errors and lack of documentation, deciphering the codes on the tapes was tedious and time consuming.

6.3 Searching via the UDC Notation

Extensive search capabilities may be employed on a mechanized data base that uses a notation such as that adopted by the UDC. One of the most effective of these is that which makes use of the hierarchical arrangement of the code. This would allow for the association of terms at: a) the same level of specificity; b) at a higher level of generality; c) to a greater degree of specificity and d) at a comparable and adjacent level of specificity.

Example:

a) same level:

556.55	lakes reservoirs ponds limnology

b) at a higher level of generality

556.5	surface water hydrology land hydrology

or even higher

556	general hydrology hydrosphere

c) to a higher degree of specificity

556.555	lake regimes

d) at an equivalent and adjacent level

556.555.2	level
556.555.3	inflow and outflow

Searching techniques could be performed on the following basis. If there are x characters in the query term, a search program will consider it a 'hit' if there is a match of the first x characters in the descriptor term of the document. For example, a request for documents on the properties of water would be encoded 556.11. This would be sufficient to retrieve a paper on the odour of water indexed by 556.114.44.

Search techniques may also take advantage of the form and context facets used in the UDC. For example only documents dealing with Alberta, (712.3), during the year "1967", and written in French, =40, on river water pollution, 556.535.8, are wanted. Once the desired facets have been established a search of all the UDC document indexes can be made. A document would then be considered a 'hit' if its UDC number matched the search criteria. To extend this even further more complex boolean logic could be included or instead, weighted boolean searches [16].

More sophisticated searches could be devised that would automatically increase the level of generality being searched for. For example, if a document were being scanned to meet the above specified search criteria and the facet, Alberta, (712.3), was not found then the search program could automatically look for the next higher level of generality.

In this case (712) which is Prairie Provinces. If weighted boolean logic were being used the weights assigned to terms could be automatically decremented as one went up the hierarchy. The sum of the weights could then also be used to indicate the relevance of a document. It should be noted that a similar extension from specific to general is possible within the Medlars system [17].

It has already been shown that the UDC can perform adequately in a computerized retrieval system by making use of simple search procedures but it was not designed with this in mind. Therefore, although it is adequate in its present form it could be improved upon. For example, some of the symbols presently used in the UDC, e.g. the colon, are too general and should be replaced by a set of more precise indicators. This change would help by more clearly defining relationships between facets significant to a document.

6.4 Testing of the UDC

6.4.1 Introductory Comments

It has been shown that the UDC is an effective tool for the classification and retrieval of information in a computer environment [15]. However, its effectiveness as an indexing language for the classification of information of a particular subject field should be determined before its use is adopted in a particular system. A detailed desciption of the method developed for determining the relevancy of the UDC as an indexing language for the classification and retrieval

of information used for the management and planning of water resources follows. This method could be applied to any system requiring a measure of the suitability of the indexing language being used or proposed.

6.4.2 General Methodology

The general methodology used for the investigation of the UDC as a tool for the classification of water resources material will be described in three steps. 1) The first step, through the manipulation of the files, is to find all subject headings in the UDC that are similar to the descriptor terms from the Water Resources Thesaurus. 2) The second step is to analyze those subject headings matched in the first step to see if they are relevant to water resources. If so, these two steps will then result in a basic concordance between the Water Resources Thesaurus and the UDC schedules; this concordance is to be used in step three. 3) Step three determines the suitability of using the UDC in conjunction with the Water Resources Thesaurus for the classification of information for water resource management. This step includes user studies. Owing to the lack of suitable facilities this step was not completed for the thesis. However, it will be done later in a working application.

Step 1

The objective of step one was to find all occurrences of water resource terms in the UDC schedules (See Figure 7). The water resource vocabulary consisted of all main terms and use references from the WRSIC thesaurus and the WPO word list. The UDC schedules consisted of the UDC English Language Master File and the recent updates to the hydrology section.

Because of the large size of the files being dealt with, the unpredictability of these files due to the occurrence of errors and poor documentation, and the necessary analysis of provisional results several programs had to be written in a methodical manner (See Figure 8). These programs were written to a) convert the files into a form suitable for automatic comparison purposes, b) compare UDC words to thesaurus terms, taking into consideration such factors as relative positions of words, and c) arrange and print the file of matched thesaurus terms with UDC subject classes, resulting from (b), into ordered printed lists. These lists were then used in step two.

Before programs could be written, a filing structure for the UDC and WRSIC files had to be decided upon. The files had to be put in a form that would allow a binary search to be carried out on the water resource terms. This meant that all water resource terms had to be placed in a table and structured in a fixed format. The way in which

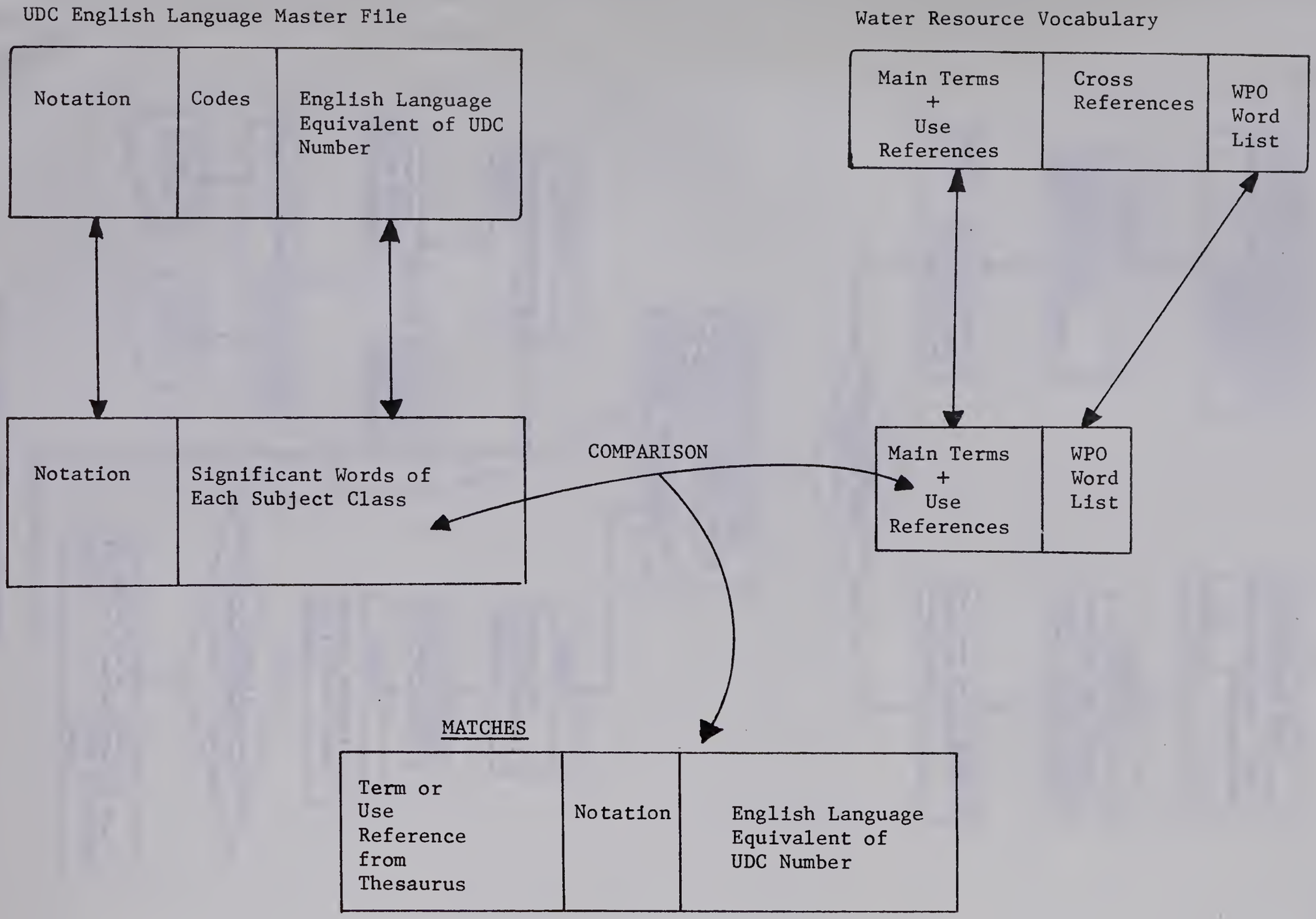


Fig. 7: Matching of UDC Vocabulary to Water Resources Vocabulary

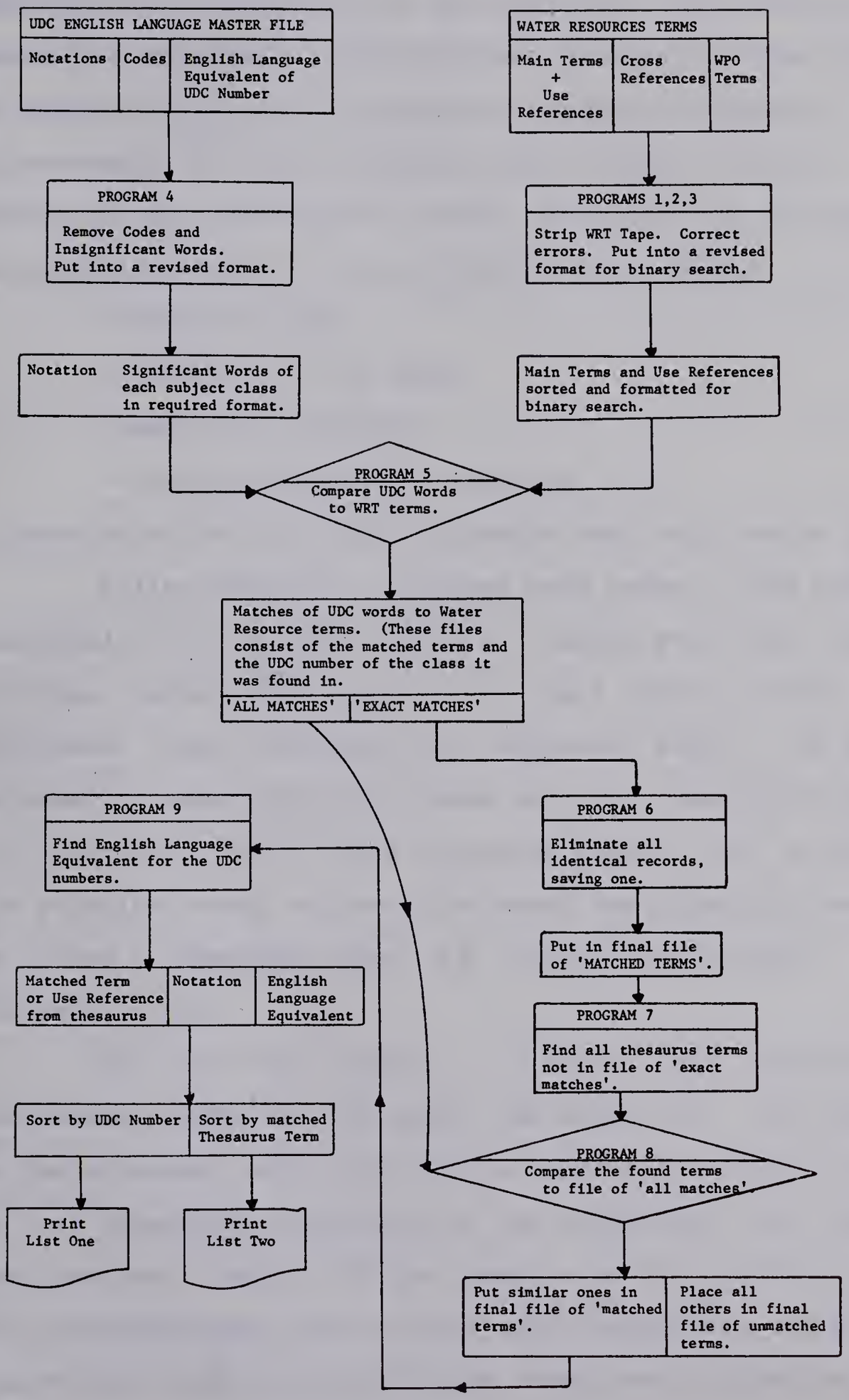


Fig. 8: Step One of Investigation Procedure

the terms were structured in the table must also allow the comparison program to find different formats of these terms. An example of the match procedure is shown in Figure 12. In this example, the term "drinking water" would be found matched to any UDC subject classes containing the English language equivalent in any of the following forms:

--drinking water

--drinking of the water

--water for drinking

--water purification, drinking, etc.

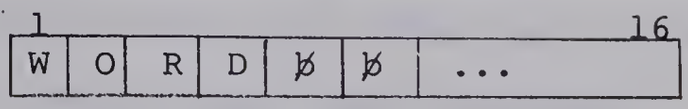
A description of the filing structure used will now be given.

Filing Structure: Because most terms in the thesaurus consisted of two or more words, two tables were built, the principal table containing all principal words, and the complement table containing all complement words. For the purposes of this thesis the first word of a descriptor term, e.g. 'drinking' in the term 'drinking water', will be called the principal word; any word following the principal word will be called a complement word, e.g. 'water' in the term 'drinking water'.

All words were placed in 16 bytes, left justified with blanks padded on the right (See Figure 9). All words in the principal table also had an additional 16 bytes tagged on that consisted of pointers to the complement table and some padding. Because of the format used all pointers in the principal table were on half word boundaries. This allowed for a much more efficient comparison program to be

written (Program 5).

Complement Word Format



Principal Word Format

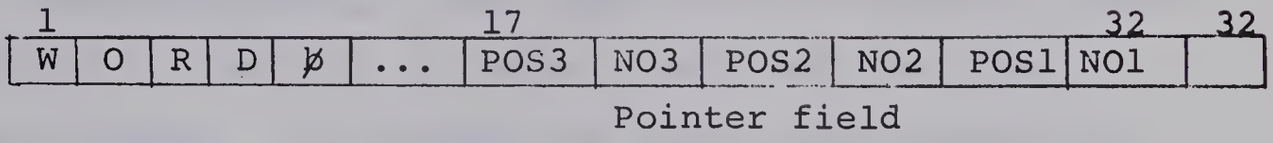


Fig. 9: Format of Terms in Principal and Complement Tables

The tables were built up sequentially in the same order as the terms were read from the thesaurus tape. The principal table was then ordered according to the collating sequence of the machine. The complement table retained the order in which it was built up. That is, any time a principal word occurred with a complement word, the complement word was placed in the complement table. Its position in the complement table was recorded as POS1, POS2, or POS3 in the pointer field of the principal word (See Figure 9). The corresponding number of complement words to occur with the principal word were then entered as NO1, NO2, or NO3 in the pointer field (See Figure 9). This means that for a given principal word, the number of complement words occurring in an uninterrupted sequence and their position in the complement table were given.

The programs used in step one were the following:

Program 1 was written to partially strip the Water Resources Thesaurus tape of cross references to provide a

file, with no duplications, of all terms used in the thesaurus. Because errors occurred in the format of the tape, the resulting file of thesaurus terms was checked for keypunching errors. Any errors found were corrected.

Programs 2 and 3 then took the file resulting from program one to build the principal table and the final complement table. In building the tables and calculating pointers for the principal words, the programs would look only at the last term entered. While so doing a count of all complement words for identical principal words was kept. However, because of the collating sequence used in the thesaurus, identical principal words, 'acid' in Figure 10, would not always appear one after the other. This meant that all complement words for one principal word would not necessarily emerge in an unbroken sequence in the complement table. Because of this several pointers were needed (See Figure 10).

Core space was not a major constraint; both the principal table and the complement table were kept in core for the comparison program. If core space had been a major factor the principal table could have been cut down to half its size. In order to do so only 12 bytes would have been allowed for the principal word and 4 bytes for one set of pointers. However, before building the tables it would have been necessary to sort the thesaurus terms from which the tables were to be built according to the collating sequence

used by the machine. This would assure that all complement words for a given principal word would appear in an unbroken sequence in the complement table thus requiring only one set of pointers. It should be noted that both the size of the word and its pointers must be kept to a power of two in order to permit an efficient binary search.

Program 4 then converted the UDC English Language Master File to a form suitable for comparison with the reformatted water resource terms. All unnecessary codes, pointers, etc., and insignificant words, such as the, and, which, etc., were stripped from each subject class. All significant words were put in a 16 byte fixed file, left justified with blank padding on the right. Figure 11 gives an example of a reformatted subject class. The file resulting from Program 4, a reformatted UDC file, contained approximately 650,000 words.

Program 5 then compared all words from the reformatted UDC file to the water resource terms. Two separate files, one file for 'exact matches' and the other file for 'all matches' were created by analyzing all subject headings in the UDC English Language Master File. The following method was used for the analysis (See Figure 12).

A subject heading from the reformatted UDC file was read into core. For each word occurring in the subject heading a binary search of the principal table was made to see if the word occurred there as well. As stated, both the

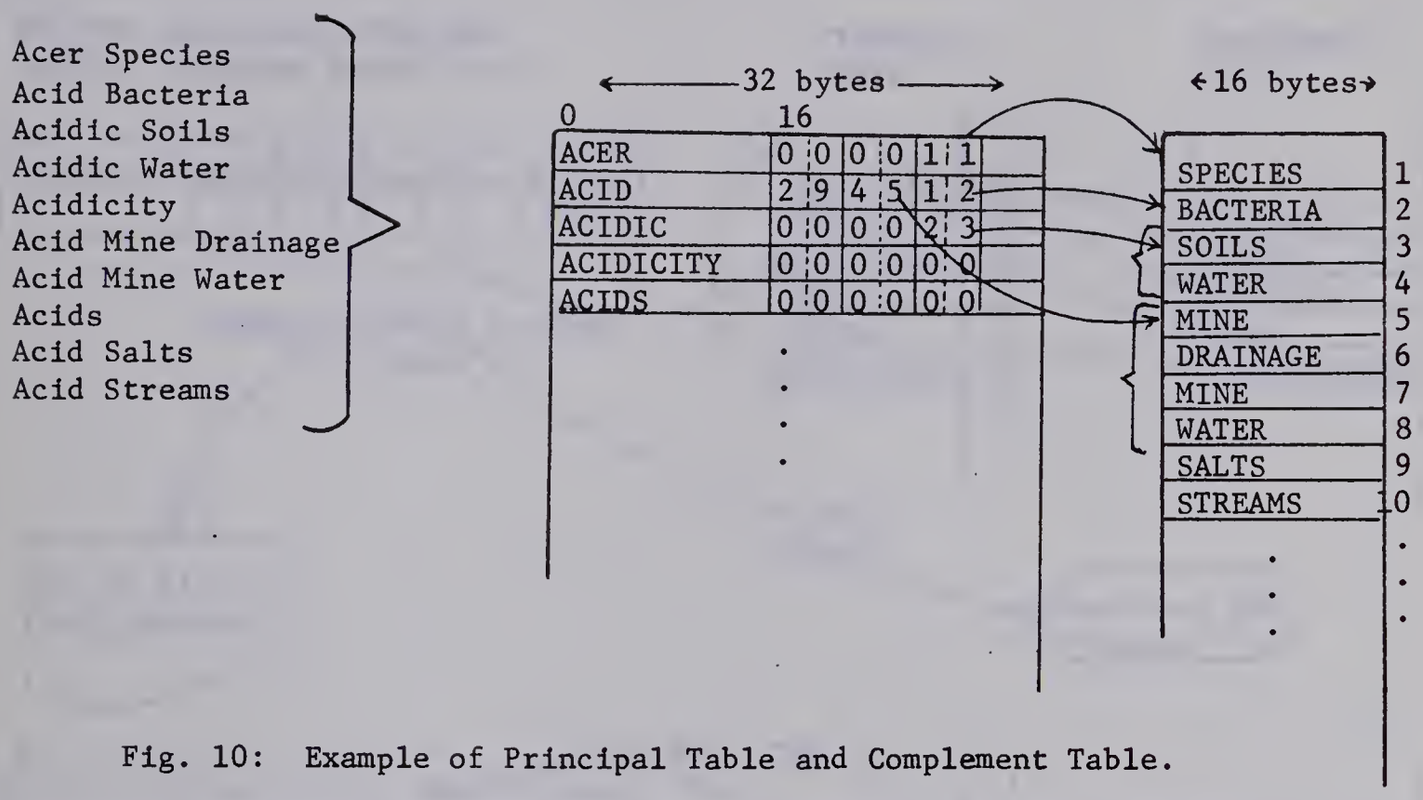


Fig. 10: Example of Principal Table and Complement Table.

Example of Subject Class from UDC English Language Master File

628Q19 ENFU--02001073$POLLUTION OF WATER SOURCES AND SUPPLY: CAUSES, PREVENTION AND REMOVAL.

Reformatted by Program 4.

628Q19	POLLUTION	WATER	SOURCES	SUPPLY	CAUSES	PREVENTION	REMOVAL
	16 bytes	16 bytes	16 bytes	16 bytes	16 bytes	16 bytes	16 bytes

Fig. 11: Example of Reformatted UDC Subject Class

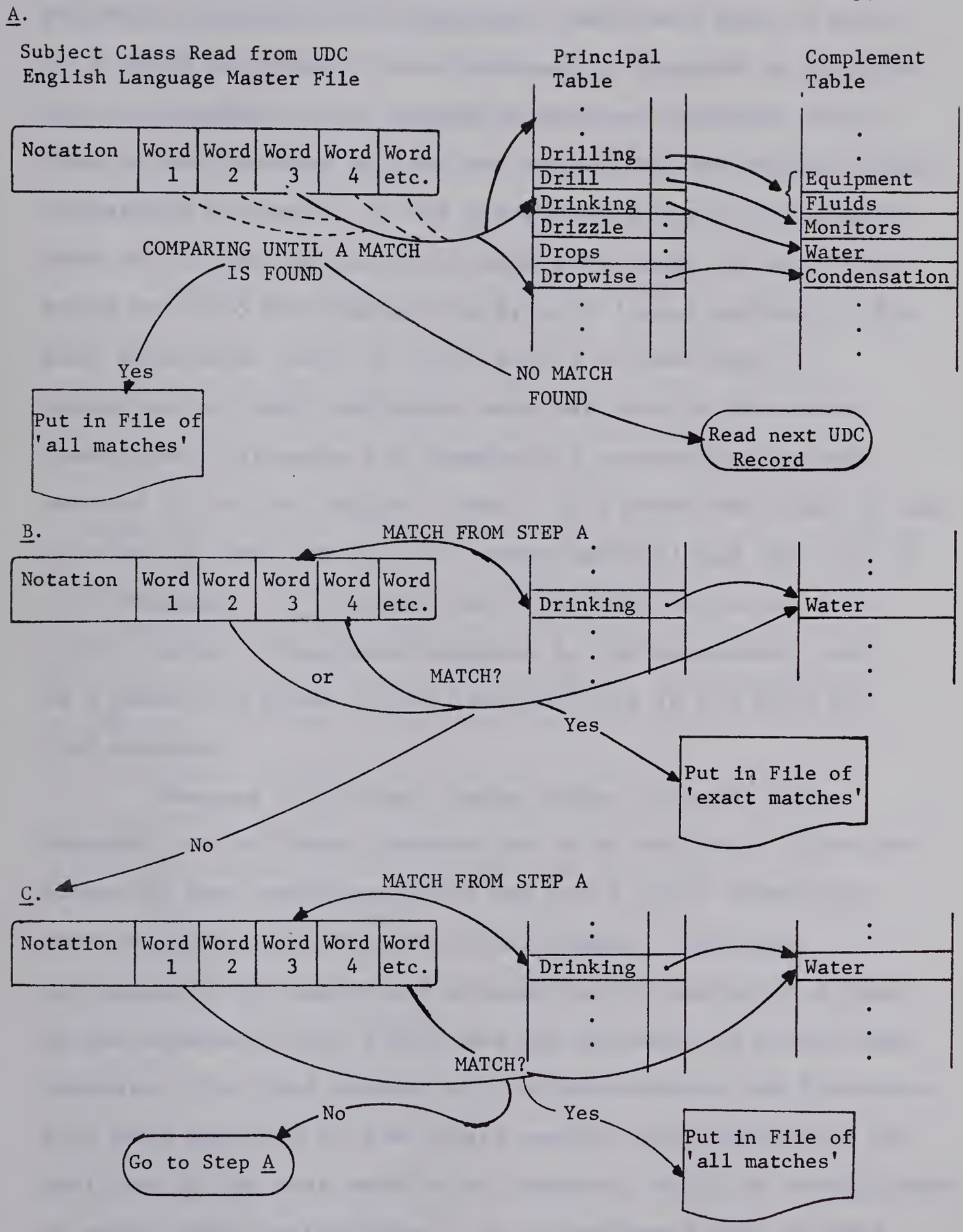


Fig. 12: Matching Procedure

principal table and the complement table were kept in core. If a match was found it was immediately recorded in the file of 'all matches'. All records of matches consisted of the term or word matched and the UDC number for the subject class containing the term. If the matched word had no complement, that is, all of its pointer fields were equal to zero, the match was also recorded in the file of 'exact matches'. For each complement found to occur with a matched word, a comparison of that complement word was made to the words immediately following and immediately preceeding the word matched in the UDC subject class. If a match was found it was recorded in both the file of 'exact matches' and the file of 'all matches'. If no match was found then all other words in the subject class were compared to the complement word. If a match was found it was recorded only in the file of 'all matches'.

Because of the very large number of words being compared, an efficient program had to be written. Since the amount of core space available was not a major constraint, speed was the major measure of efficiency. This was accomplished by keeping all entries in the tables on a power of two boundary (e.g. $2^5=32$) and all pointers on a half word boundary. For this reason, all multiplications and divisions, that were required in the binary search, for calculating the position of the next word to be compared, could be accomplished by using shift instructions. It is estimated that in this application the shift instructions lowered the time required

for the binary search by at least a factor of four.

Approximately five minutes were required by this program, when executed on an IBM 360/67, to compare the reformatted UDC file to the reformatted Water Resources Thesaurus.

Program 6 was written to eliminate all redundant matches in the files of matched thesaurus terms. Redundant matches were caused by similar sections, from different editions of the UDC schedules, being included in the UDC English Language Master File. For example, all the classes in the Abridged Edition are also included in the Full Edition. When the class 628, Public Health Engineering, is included from both editions, all entries from the Abridged Edition would be duplicated by the Full Edition. Although this did lead to duplications in matches it also resulted in matches that would not normally have been found. This is because some editions are more verbose when describing a subject class or the terminology used is more up to date than others.

Program 7 then created a file of thesaurus terms not found in the file of exact matches. This was done by comparing the file of unique thesaurus terms resulting from program one, to the file resulting from program six.

Program 8 then compared all those terms resulting from program seven to the file of 'all matches'. All terms found to occur in the file of 'all matches' were then placed in the final file of matched terms with the UDC number for the class they were found in. All other terms resulted in the final file of unmatched thesaurus terms. Further details of this stage of the matching process are to be found in reference [3].

Program 9 then found the English Language Equivalent for all UDC numbers in the final file of matched terms. This was done by first sorting the final file of matched terms by UDC number and then comparing it to the UDC English Language Master File. Because the UDC Master File was already sorted by UDC number, the actual comparison was done by stepping through both files while comparing them to one another. Whenever a match occurred the English Language Equivalent was stripped from the Master File. The resulting file consisted of a thesaurus term, the UDC number for the subject class the term was found in, and the English language equivalent for that class. Because of the procedure used by program nine the final file emerged sorted by UDC number. This file was then printed as List 1. The file was then sorted by thesaurus terms and printed as List 2. Each printout consists of the word that was found, the UDC number for the subject class it was found in, and the UDC English language equivalent for that class. Samples from these two lists are shown in Table 3.

TABLE 3

EXAMPLE OF MATCHES SORTED BY: A) UDC NOTATION, and B) THESAURUS ENTRY

A. List 1

Thesaurus Entry	Notation	English Language Equivalent
PORTLAND CEMENTS	666.94	CEMENTS
CEMENTS	666.94	CEMENTS
PORTLAND CEMENTS	666.942	PORTLAND CEMENTS
CEMENTS	666.942	PORTLAND CEMENTS
CEMENTS	666.948	ALUMINOUS CEMENTS
MATERIALS	666.95	SULFATE, TRASS CEMENTS
CEMENTS	666.95	SULFATE, TRASS CEMENTS
MATERIALS	666.96	HARDSETTING CEMENT COMPOSITIONS
MORTAR	666.97	CONCRETE
UNDERWATER	666.97.0.15	UNDERWATER SETTING
TREATMENT	666.97.0.3	PREPARATION. PROCESSING. TREATMENT
MACHINERY	666.97.0.5	PLANT, MACHINERY AND EQUIPMENT

B. List 2

Thesaurus Entry	Notation	English Language Equivalent
CEMENTS	666.94	CEMENTS
CEMENTS	666.942	PORTLAND CEMENT
CEMENTS	666.948	ALUMINOUS CEMENTS
CEMENTS	666.95	SULFATE, TRASS CEMENTS
CEMENTS	691.5	BINDING AND BEDDING MATERIALS
CEMENTS	691.54	CEMENTS
CEMENTS	691.541	ROMAN CEMENTS
CEMENTS	691.544	OTHER CEMENTS
CENSUS	312	DEMOGRAPHY. POPULATION. STATIS...
CENSUS	351.755.3	CENSUS REGISTERS. C.F. 312
CENSUS	656.0.21	NUMBER AND FREQUENCY OF TRANSPORT IN GENERAL. TRAFFIC SURVEYS.
CENTRIFUGAL PUMPS	621.67	CENTRIFUGAL PUMPS. AXIAL FLOW. AND TURBO-PUMPS.
CENTRIFUGAL PUMPS	621.671	CENTRIFUGAL PUMPS IN GENERAL
CENTRIFUGAL PUMPS	621.671.1	OPEN CENTRIFUGAL PUMPS
CENTRIFUGAL PUMPS	621.671.2	CLOSED CENTRIFUGAL PUMPS
CENTRIFUGAL PUMPS	621.671.5	CENTRIFUGAL PUMPS WITH INVOLUTE

C. Thesaurus Form

Cements
  NT Portland Cements
  RT Adhesives
  -Aggregates
  Alkali-Aggregate Reaction
  Asphalt
  Cement Grouting
  -Clays
  (etc. as in Table 4)

Step 2

Step 2 consisted of analyzing by hand those printed lists resulting from step one. These lists were used to 'link' thesaurus terms to subject classes. Since many of the terms occurred only a few times, it was not difficult to see, by using list 2 only, if the matched class would serve as a relevant 'entry point' or 'link' for that term. It was more difficult to identify an entry point for terms that occurred frequently. Many terms such as air, acids, buildings, matched a thousand times or more. Such a large number of matches is generally caused by a term, such as acids, appearing in a 'roof class' and many of its subclasses as well.

No set procedure or algorithm could be derived for linking the thesaurus terms to subject classes because each matched term had to be evaluated separately. In spite of difficulties, certain basic steps that were used for most terms have been established. The following is a description of these:

Take a unique term in List 2. For example, from the subset of List 2 illustrated in Table 3(b) the first term to be looked at would be "cements". Then the thesaurus would be checked to see the relationships of the term in the thesaurus (Table 3(c)). In an on-line implementation the relationship would be displayed automatically by accessing the thesaurus. As stated the joint work with Mr. Fred Alber will result in an on-line water resources thesaurus. The order of the UDC

numbers for the classes in which the word occurred was then checked to see if some type of structure was evident. From the structure a possible 'entry point' or 'link' to the schedules may emerge. From List 2 in Table 3 the possible major entry points given to the term "cements" were 666.9, since four of the matches are subsets of this class, and 691.5, again because four of the matches are subsets of this class. Once possible entry points have been determined a check is made of the Engish language equivalent for that class and its surrounding classes to see if it is relevant or not. In order to help determine relevancy the schedules and/or List 1, sorted by UDC number, must be consulted.

Once all non-relevant entry points are eliminated it is necessary to decide on the validity of those left. In this particular use this is done by keeping those classes which specifically serve the classification of water resources information and discarding those that do not. It must also be decided whether, for the use envisaged, the entry point is too high or too low in the structure of the class being entered.

With this procedure entry points for the term "cements" would be 666.9 and 691.5; entry point for "census" would be 312; entry point for "centrifugal pumps" would be at 621.671 because an entry at 621.67 covers axial-flow and turbo-pumps also. It should be realized that List 1 in Table 3 is a subset of the complete list that was used in making decisions such as those above. Note that this step results

in an initial concordance between the UDC and the Water Resources Thesaurus. An example of the form of this concordance is given in Table 4. Step two was strictly an intellectual task. Human procedures, too complex to program, were required. Definite logical patterns did occur for some subject classes that were found to match a given water resource term. However, it often occurred that matched subject classes did not fit a common pattern and it was one of these that was found to fit the context of the water resource term.

Step 3

The concordance of the Water Resources Thesaurus with the UDC resulting from the previous two steps was to be tested against the Planning Division Data Base and the Geology Document File. This modified thesaurus is to be used both as a guide to indexing documents, and as a guide to the automatic retrieval of documents via UDC number and/or keywords.

User tests of the system were not carried out for this thesis due to lack of time. However, these tests will be administered in a working environment where sufficient time and resources are available.

EXAMPLE OF CONCORDANCE

CATCH CROPS
  RT (CONTINUED)
  DROUGHT TOLERANCE
  DRY FARMING
  FARM MANAGEMENT
  -FREEZING
  LAND USE
  STORMS
  WATER INJURY
  WINTERKILLING

CATCHMENT BASINS
  USE WATERSHEDS (BASINS)

CATFISHES
  UF BLINDCATS
  BLUE CATFISH
  HEADWATER CATFISH
  ICTALURUS CATUS
  ICTALURUS FURCATUS
  ICTALURUS LUPUS
  NOTURUS SPECIES
  WHITE CATFISH
  BT -ANIMALS
  -AQUATIC ANIMALS
  -AQUATIC LIFE
  -FISH
  -FRESHWATER FISH
  -PAN FISH
  -WILDLIFE
  NT BULLHEADS
  CHANNEL CATFISH
  MADTOMS
  RT ROUGH FISH

CATHODES 621.3.032
  BT -ELECTRODES
  -EQUIPMENT
  RT CORROSION CONTROL
  ELECTROCHEMISTRY
  ELECTROLYSIS

CATHODIC PROTECTION 620.197.5
  RT -COATINGS
  -CORROSION
  -LININGS
  PITTING (CORROSION)

CATION ADSORPTION
  BT -ADSORPTION
  RT ANION ADSORPTION
  -CLAY MINERALS
  IONS

CATION EXCHANGE
  BT -ION EXCHANGE
  -SEPARATION TECHNIQUES
  RT ANION ADSORPTION
  ANION EXCHANGE
  -DEMINERALIZATION
  PERMSELECTIVE MEMBRANES

CATTAILS
  BT -AMPHIBIOUS PLANTS
  -AQUATIC LIFE
  -AQUATIC PLANTS
  -MONOCOTS
  -PLANTS
  -ROOTED AQUATIC PLANTS
  RT -AQUATIC WEEDS
  RIPARIAN PLANTS

CATTLE 636.2
  BT -ANIMALS
  -DOMESTIC ANIMALS
  -LIVESTOCK
  -MAMMALS
  -RUMINANTS

CAVEFISHES
  BT -ANIMALS
  -AQUATIC ANIMALS
  -AQUATIC LIFE
  -FISH
  -FRESHWATER FISH
  -WILDLIFE
  RT CAVES

CAVERNS
  USE CAVES

CAVES 551.44
  UE CAVERNS
  RT CAVEFISHES
  KARST
  LIMESTONES
  PORES
  SINKS
  TRAVERTINE

CAVITATION 532.528
  RT AERATION 620.193.16
  AIR DEMAND
  BUBBLES
  -CONDUITS
  -CORROSION
  -EROSION
  -FLOW
  FLOW SEPARATION
  -FLUMES
  -HYDRAULICS
  -HYDRAULIC STRUCTURES
  HYDROFOILS
  IMPELLERS
  -IRRIGATION WATER
  NEGATIVE PRESSURE
  -PRESSURE
  REAERATION
  SCOUR
  VORTICES

CELERITY
  RT -RATES
  STANDING WAVES
  -WAVES (WATER)

CELLS (BIOLOGICAL)
  USE CYTOLOGICAL STUDIES

CELLULOSE 547.458.8
  BT -CARBOHYDRATES 661.728
  -ORGANIC COMPOUNDS
  RT FIBERS (PLANT) 676.16
  LIGNINS
  LUMBER
  PULP WASTES
  RESINS
  -VASCULAR TISSUES
  WOOD WASTES

CEMENT GROUTING
  BT -GROUTING
  RT -CEMENTS
  MORTAR

CEMENTS 666.9
  NT PORTLAND CEMENTS 691.54
  RT ADHESIVES
  -AGGREGATES
  ALKALI-AGGREGATE REACTION
  ASPHALT
  CEMENT GROUTING
  -CLAYS
  -CONCRETE ADDITIVES
  CONCRETE MIXES
  -CONCRETES
  -CONCRETE TECHNOLOGY
  -CONSTRUCTION MATERIALS
  DIATOMACEOUS EARTH
  -GROUTING
  -LININGS
  MASONRY
  MORTAR
  PAINTS
  POZZOLANS
  SLURRIES
  SOIL CEMENT

CENOZOIC ERA (118|119)
  BT -GEOLOGIC TIME
  NT PLEISTOCENE EPOCH 551.7
  -QUATERNARY PERIOD
  RECENT EPOCH
  TERTIARY PERIOD

CENSUS 312
  UE INVENTORYING
  NT CREEL CENSUS
  RT -DATA COLLECTIONS
  ESTIMATING
  FUTURE PLANNING (PROJECTED)
  HISTORY
  -INVESTIGATIONS
  MARKING TECHNIQUES
  -MEASUREMENT
  -PLANNING
  -POPULATION
  -SAMPLING
  STANDING CROP
  STATISTICS
  -SURVEYS
  -WILDLIFE

CENTRAL U.S.
  UE MIDEAST U.S.
  MIDWEST U.S.
  BT -GEOGRAPHICAL REGIONS
  -REGIONS
  NT ARKANSAS
  COLORADO
  -CORN BELT
  -GREAT PLAINS
  ILLINOIS
  INDIANA
  IOWA
  KANSAS
  KENTUCKY
  MICHIGAN
  MINNESOTA
  MISSOURI
  MONTANA
  NEBRASKA
  NEW MEXICO
  NORTH DAKOTA
  OHIO
  OKLAHOMA
  SOUTH DAKOTA
  TENNESSEE
  TEXAS
  WISCONSIN
  WYOMING
  RT -GREAT LAKES REGION
  -ROCKY MOUNTAIN REGION
  -SOUTHEAST U.S.
  -SOUTHWEST U.S.

CENTRAL VALLEY PROJECT
  RT -CALIFORNIA
  MULTIPLE-PURPOSE PROJECTS

CENTRARCHUS SPECIES
  USE SUNFISHES

CENTRIFUGAL PUMPS 621.67
  BT -EQUIPMENT
  -PUMPS
  RT -TURBINES

CENTRIFUGATION
  RT -ANALYTICAL TECHNIQUES
  FILTRATION
  -WATER PURIFICATION

CERATOPOGONIDS
  USE MIDGES

CERCARIAE
  USE TREMATODES

CEREAL CROPS 633.1
  BT -AGRONOMIC CROPS
  -CROPS
  -FIELD CROPS
  -GRASSES
  -MONOCOTS
  -PLANTS
  NT BARLEY
  CORN (FIELD)
  OATS
  RICE
  WHEAT
  RT GRAINS (CROPS)
  PLANT GROUPINGS
  -SOIL MANAGEMENT
  SORGHUM
  WILD RICE

CESIUM 546.36
  BT -ALKALI METALS 553.633
  -INORGANIC COMPOUNDS
  -METALS

CESSPOOLS 628.43
  RT DISPOSAL 696.138
  DOMESTIC WASTES
  LEACHING
  SEPTIC TANKS
  SEWAGE DISPOSAL
  SUMPS
  -WASTE DISPOSAL
  -WASTE WATER DISPOSAL
  -WASTE WATER DISPOSAL

CHAMAECYPARIS SPECIES
  USE WHITE-CEDAR TREES

CHANGE ORDERS
  BT -ADMINISTRATION
  -CONTRACT ADMINISTRATION
  -CONTRACTS
  RT CLAIMS (CONTRACTS)
  -CONTRACTS
  -LEGAL ASPECTS
  NEGOTIATIONS

CHANGULA HYEMALIS
  USE OLDSQUAW DUCK

CHANNEL CATFISH
  UF ICTALURUS PUNCTATUS
  BT -ANIMALS
  -AQUATIC ANIMALS
  -AQUATIC LIFE
  -CATFISHES
  -FISH
  -FRESHWATER FISH
  -PAN FISH
  -WILDLIFE

CHANNEL EROSION
  BT -EROSION
  NT GULLY EROSION
  RT BANK EROSION
  -BANKS
  -BEDS
  CHANNEL IMPROVEMENT
  CHANNEL MORPHOLOGY
  -CHANNELS
  HILL EROSION
  RIVER TRAINING
  SALTATION
  SEDIMENT CONTROL
  SOIL EROSION
  STREAM EROSION

UE used for, BT broader term, NT narrower term; RT - related term; a dash (-) preceding a term means there are narrower terms

CHAPTER VII

STUDY OF THE LIBRARY OF CONGRESS CLASSIFICATION SCHEME

The Library of Congress classification scheme was developed for the United States Library of Congress at the turn of this century. It is intended to fit the Library's collections and services as precisely as possible without reference to outside needs or influence [18] [19] [20].

The scheme is partitioned with the objective of securing well defined areas corresponding to the concepts by which separate fields are taught and expounded within each area to provide an orderly arrangement, to make direct access for scholars and qualified students and to be helpful to the staff of the Library of Congress. A certain amount of independence applies to common elements which are treated differently in different parts.

The notation applied to the scheme is fixed and the different symbols used fall into a clear pattern. The main classes are denoted by a capital letter and in most a second capital to denote major sections. Arabic numberals are then used to denote sections. These are used integrally from 1 to 9999 if necessary, with gaps left liberally to accommodate new topics as they arise. The notation does not dictate order; if a new topic has to be inserted where no gaps exist, a decimal point is used for further subdivision.

Further arrangement is often alphabetical with some use of Cutter numbers at certain points. These numbers consist of a capital letter followed by one or two digits to give a symbol for the name of a topic or an author.

Each main class has its own index with the exception of P and some parts of A. However, no overall index exists for the whole schedules as is the case for the UDC. When a book is to be classified the correct main class must first be selected before the index can be consulted. In some cases the index to one class may contain cross references to another but these cross references are exceptions rather than the rule.

7.1 Outline of the LC Classification

The outline of the scheme is dictated by the organization of the Library of Congress rather than by theoretical considerations. As a result of being matched to the needs of the sections of the individual LC collections, there are no common facets in the schedules. This leads to very bulky schedules, which at present consist of 21 main classes occupying 6000 pages with Literature and Language being about 1/3 of this. These schedules are not available in machine readable form.

The main classes are as follows:

A Generalia. Some literal mnemonics are included, eg. AE Encyclopedias. AZ is used for History of the sciences in general, Scholarship, Learning.

B-BJ Philosophy. This includes Psychology, Ethics, and Etiquette.

BL-BX Religion.

C Auxiliary Sciences of History. Archeology CC, and Numismatics CJ and Collective Biography CT are included.

D History: General and Old World. DA is Great Britain. Other European countries are in approximately alphabetical order.

E-F History: America.

G Geography. Maps and atlases are included. Most branches of Geography are in this schedule, also related topics such as Anthropology GN, Folklore GR and Recreation GV.

H Social Sciences. Economics occupies HB-HJ. Socialism, Marxism and Communism are in HX. Criminology is in HV.

J Political Science.

K Law. Not yet published.

L Education. Much of the schedule is simply a listing of educational establishments under country.

M Music.

N Fine Arts.

P Language and Literature.

Q Science. There is no synthesis in this class at all.

Very little systematic arrangement except by alphabetical order.

R Medicine. Primary division is by medical discipline, Surgery RD.

S Agriculture. Crop subordinate to pest at SB608. SB975 ends with SB987 General Works; there is no provision for the treatment of a particular pest by a particular method.

T Technology. Alphabetical arrangement is widely used. There is little provision for composite subjects.

U Military Science.

V Naval Science.

Z Bibliography.

7.2 Retrieving from LC

Theoretically, the main tool used for retrieving from an LC classified manual data base (i.e. an LC classified library) is the LC subject heading catalogue as used at the installation. This catalogue should correspond to the subject headings on the LC card catalogues, i.e. should represent the complete entry vocabulary. Synonyms are given for the headings used and in some cases scope notes and cross references. Natural language is used almost without exception.

In these catalogues, certain categories of heading are not incorporated. These include persons, family names,

corporate bodies, structures, ships, religious bodies, and mythological characters. They may, however, appear in the list if they are used as examples under other headings. However, such authority lists do not always correspond to the subject headings or are not made available to the user public. The user's main entry point is the subject heading card in the card catalogue which again should correspond to the subject heading listed on each bibliographic description card.

The MARC tapes consist of card images representing a data base, such as a card catalogue, available to users. In this study, therefore, the MARC tapes were used both because they provided a means of linking classifications to documents and to subject words and because these subject words represented user access points. A case might be made that the LC subject headings tape should be used and cross checked for classification linkage and for a complete study this should be done. However, it was not possible to do this within the limits of this thesis.

7.3 Testing of the LC

The approach used for the machine testing of the classification was to take all the main terms and use references of the Water Resources Thesaurus, that is, the same list used in the UDC investigation, and to compare the terms with the subject headings stripped from the MARC tapes. As stated, the MARC tapes were used because they were

considered to represent the current vocabulary as expressed through the subject headings. All subject headings matched to the thesaurus entry were stripped off with their associated LC number.

Two basic steps were involved in the stripping procedure. The first step created a file of all subject headings, occurring on the MARC tapes, with their associated LC number. The file was sorted according to subject heading entry (See Table 6). The file was created by running against the MARC tapes programs that were previously developed at the University of Alberta. One of the options provided by these multi-purpose programs is to strip the subject headings from the MARC tapes. These programs were written by V. Shapiro, D. Walker, and F. Appleyard [20]. It should be noted that there were approximately 26000 MARC records and 29000 LC subject headings.

The second step was to write a program that would automatically compare the sorted LC MARC subject heading file, resulting from the above step, with the Water Resources Thesaurus file which had resulted from the UDC investigation. This comparison was done by automatically stepping through the thesaurus file and the subject heading file while comparing them with one another. When a match was found the thesaurus term was written out with the matching subject headings listed below it with the associated LC numbers (See Table 5). In this part of the investigation the author was assisted by P. Laffin and G. Ewing.

TABLE 5

STRIPPING OF LC CALL NUMBER AND
SUBJECT HEADINGS

CURRENCY QUESTION CHINA.	HG4572 C75 1968
CURRICULUM ENRICHMENT.	LC3993 .R58
CURVES JUVENILE LITERATURE.	QA484 .R38
CURVES.	QA484 .R38
CYTOLOGY.	QH581 .B77
CZECHS IN THE UNITED STATES.	E184.B67 C29 1969
DAIRY PRODUCTS ADDRESSES, ESSAYS, LECTURES.	QP751
DAMS CALIFORNIA.	TC557.C2 A45
DEACONS.	BV680 .T5
DEACONS.	BX1912
DEAF BALTIMORE.	HV2561.M3 F79

TABLE 6

EXAMPLE OF LC MATCHES

W OIL FIELDS

OIL FIELD FLOODING ADDRESSES, ESSAYS, LECTURES	TN870 .p449 1967AA
OIL FIELD FLOODING WATER-SUPPLY	TD224.T4 A333 NO. 44
OIL FIELD FLOODING	HD9564 .C23 NO. 67
OIL FIELDS ALBERTA	HD9564 .C23 NO. 67
OIL FIELDS PRODUCTION METHODS ADDRESSES, LEC...	TN870 .P449 1967AA
OIL FIELDS TEXAS	TD224.T4 A333 NO.44
OIL FIELDS U.S.	TN872.A5 W38

W OUTDOOR RECREATION

OUTDOOR RECREATION	TC424.C2 A62 1968
OUTDOOR RECREATION CALIFORMIA OROVILLE REGION	GV54.C2 S635
OUTDOOR RECREATION ECONOMICS ASPECTS CALIFORNIA	J87 .C2 1965-1967
OUTDOOR RECREATION ECONOMICS ASPECTS U.S.	SB481 .B4
OUTDOOR RECREATION FLORIDA EVERGLADES	S972.F58 C4
OUTDOOR RECREATION GT. BRITAIN	HD951 .A1W9
OUTDOOR RECREATION GT. BRITAIN	GV75 .B85 1967
OUTDOOR RECREATION LAW & LEGISLATION. VIRGINIA	
OUTDOOR RECREATION LAW & LEGISLATION. U.S.	KF26.1542 1969
OUTDOOR RECREATION MICHIGAN	QE125 .A37 NO. 3
OUTDOOR RECREATION MISSOURI VALLEY	GV54.M8 U5
OUTDOOR RECREATION NEVADA FINANCE	JK8501 .N45 NO.3
OUTDOOR RECREATION NEW HAMPSHIRE	GV54.N4 A2 NO.13
OUTDOOR RECREATION NEW MEXICO	GV54.N6 A5 1968
OUTDOOR RECREATION OREGON	GV54.07 A35 1967
OUTDOOR RECREATION TENNESSEE FINANCE	JK5274 .A27 1968,B13
OUTDOOR RECREATION U.S.	GV53 .A46
OUTDOOR RECREATION VERMONT	GV54.V5 A43
OUTDOOR RECREATION WASHINGTON (STATE)	GV54.W2 A4
OUTDOOR RECREATION WASHINGTON (STATE)	GV54.W2 A5 1968
OUTDOOR RECREATION WISCONSIN	S932.W6 A45
OUTDOOR RECREATION	GV182.2 .D3
OUTDOOR RECREATION	GV182.2 .P7

The comparison technique employed took only the first x-1 characters of the term, x being the number of characters in the term; then these are compared with the first x-1 characters of the subject headings. This technique accounts for most plurals that occur in either file, but it also leads to some strange matches, such as CHERT from the thesaurus being matched to CHEROKEE from MARC. Further programming could have eliminated the majority of these false hits by providing for the checking of the last character of the thesaurus term and then taking precautionary steps if the character is an 'a', an 'l', or an 's'. The extra time to produce this program was shown to be unnecessary since the false hits were few and obvious and could easily be found by scanning the printed lists.

The file resulting from the comparison was then studied. A sub-set of this file is given in Table 5. At first glance, there seems to be no definite 'entry points' from the thesaurus to the schedules. For example, from Table 5, it is apparent that the term "outdoor recreation" leads to several main classes in the LC schedules, GV54, GV182, and TC424. This means that a descriptor from the thesaurus could lead to several subject classes all with very distinct LC numbers with no interconnection between them. No obvious matches appeared as did with the UDC comparisons.

However, these results are not conclusive. As stated, the LC investigation will be carried further. More

sophisticated matching procedures need to be devised and use should be made of the LC subject heading tapes. Nevertheless, if we postulate that the MARC subject headings constitute the indexing and searching vocabulary, the lack of pattern in the matches is significant. This fact helps support a general conclusion, reached by studying the structure of the LC classification, that it would not lend itself readily to on-line automatic retrieval methods.

CHAPTER VIII

RESULTS OF CLASSIFICATION TESTING

8.1 General Results

Testing of a classification scheme through the use of the method described in this thesis has several distinct advantages. Some of these are the following:

1. A better understanding of the classification is gained. This is because concepts within the classification are better illustrated; hidden concepts may emerge; poorly defined concepts become clear.

2. The development of ideas and categories within the classification is illustrated. This reflects on the organization of the framework or structure of the classification and the degree to which this organization is adhered to. This also indicates the suitability or unsuitability of the classification for automated retrieval methods.

3. Those classes that are more important and more useful for classifying information in a given subject area stand out and are made obvious. Because of this, it may be shown that only subsets of an overall classification are needed for classifying a given data base. These subsets could generally be improved upon by being expanded to more detail to suit the given needs.

4. The current state of development, or the provisions made within a classification, for a given subject area are shown.

5. A measure of the degree to which the data bases tested are interrelated is given.

6. A working concordance results giving the linkage between the classification being tested and the thesaurus from which the test vocabulary was derived. This concordance could be extended and refined through use.

It should be noted that this concordance is not final. It would have to be refined and extended to increase its effectiveness. This refinement will be done in the Water Planning and Operations Branch of the Federal Department of Environment. However, it should be made clear that in the manual development of a concordance a large part of the work is routinely taken up by clerical tasks. These are very time consuming. The procedures developed in this thesis perform the routine clerical tasks by computer and separate them from the intellectual tasks required when building a concordance.

Overall, the algorithm developed in this thesis shows that large data bases can be effectively handled to bring out their various characteristics.

8.2 Results and Conclusions of UDC Testing

The results of the UDC testing are given in Table 7. The conclusions arising from these results are as follows:

TABLE 7

RESULTS OF UDC TESTING

NUMBER OF THESAURUS TERMS	5180
NUMBER OF WPO TERMS	234
TOTAL NUMBER OF WATER RESOURCE TERMS	5414

PROGRAMMING RESULTS

THESAURUS TERMS MATCHED TO UDC ENGLISH LANGUAGE MASTER FILE	2560
THESAURUS TERMS MATCHED TO UDC ENGLISH LANGUAGE MASTER FILE AND UDC UPDATES	2594
WPO TERMS MATCHED TO UDC ENGLISH LANGUAGE MASTER FILE	58
WPO TERMS MATCHED TO UDC ENGLISH LANGUAGE MASTER FILE AND UDC UPDATES	70
WATER RESOURCE TERMS MATCHED TO UDC ENGLISH LANGUAGE MASTER FILE AND UDC UPDATES	2664 or 49.2%

RESULTS AFTER EXAMINATION OF COMPUTER MATCHES

NUMBER OF WATER RESOURCE TERMS WITH MATCHES IRRELEVANT TO WATER RESOURCES	458
TOTAL NUMBER OF WATER RESOURCE TERMS WITH MATCHES RELEVANT TO WATER RESOURCES	2206 or 40.7%

1. The vocabulary used in the parts of the UDC schedules relevant to the classification of water resources material appears to be up to date. This is indicated by the percentage of words from the Water Resources Theaurus that occur in the UDC schedules.

2. An overall set of UDC schedules would not be needed to classify water resource material. The classes that stood out as a result of the testing are given in Table 8. Rather, what would be needed is the development of a set of schedules for water resources similar in form to those used for other special collections such as the Scott Polar Library [21].

3. The UDC, combined with the thesaurus, would be an effective way of classifying water resource material.

4. In most instances, plurals are used in the UDC schedules in the same manner as they are used in the thesaurus.

5. A concordance resulting from the UDC schedules and the thesaurus would serve to indicate when sections of the schedules need updating.

6. It appears the UDC could be regarded as a metalanguage which could be used effectively for automated retrieval purposes for water resources management. This conclusion was reached from the study of the notation.

7. Using the concordance resulting from this study will allow the user to perform more effective searches on UDC mechanized data bases.

TABLE 8

UDC CLASSES THAT STOOD OUT

SMALL PARTS OF	33	ECONOMICS.
SMALL PARTS OF	34	LAW AND LEGISLATION.
SMALL PARTS OF	35	PUBLIC ADMINISTRATION. GOVERNMENT.
SMALL PARTS OF	36	SOCIAL RELIEF. WELFARE. INSURANCE.
SMALL PARTS OF	37	EDUCATION.
SMALL PARTS OF	38	COMMERCE.
	517	ANALYSIS. CALCULUS. FUNCTIONS.
	518	MATHEMATICAL MODELS.
	519	STATISTICS.
	528	GEODESY. PHOTOGRAMMETRY. CARTOGRAPHY.
	53	PHYSICS.
	54	CHEMISTRY.
	55	GEOLOGY. METEOROLOGY.
	576	CYTOLOGY. HISTOLOGY. ORGANOLOGY. EVOLUTION.
	577	GENERAL PROPERTIES OF LIFE.
	581	PLANT PHYSIOLOGY.
	582	SYSTEMATICS.
	59	ZOOLOGY.
	614	PUBLIC HEALTH AND SAFETY. ACCIDENT PROTECTION.
	615	PHARMACY.
	616	PATHOLOGY.
	62	ENGINEERING AND TECHNOLOGY.
	63	AGRICULTURE. FORESTRY. FISHERIES.
	66.0	CHEMICAL ENGINEERING.
	661	CHEMICALS.
	663.6	WATER FOR DRINKS.
	666.9	GYPSUM, LIME AND CEMENT INDUSTRIES.
SMALL PARTS OF	67/68	VARIOUS INDUSTRIES AND CRAFT.
SMALL PARTS OF	69	BUILDING INDUSTRY.
	711	PHYSICAL, REGIONAL, TOWN AND COUNTRY PLANNING.
	712	LANDSCAPE, NATURAL AND DESIGNED. PARKS.
SMALL PARTS OF	79	ENTERTAINMENT, GAMES AND SPORT.
	91-1/-9	GEOGRAPHY.

8. The concordance will also serve effectively as the heart of an on-line information storage and retrieval system and will be used in both searching and indexing.

8.3 Results and Conclusions of LC Testing

The results obtained from the LC testing are more general and less satisfactory than those from the UDC test. As the study progressed it was realized that the LC would not be very suitable in a mechanized environment mainly because of its notation and structure; the suitability in a mechanized environment was one of the criteria set down for the chosen classification. Nevertheless, it was decided to continue the study of the LC but on a less detailed basis. The further tests carried out confirmed the early suspicions of the unsuitability of LC in a mechanized environment. Results of these tests are given in Table 9. The conclusions that were made after testing the LC were as follows:

1. At present, the LC would not be suitable in a mechanized environment.

2. LC subject headings on the MARC tapes relevant to water resources are not up to date although we would suppose that current MARC tapes would represent current vocabulary. This is indicated to some extent by the very low number of water resource terms that were matched. In addition, many of the terms matched were not terms very specific to water resources but terms that are generally used in most other subject areas. Examples of these terms

TABLE 9

RESULTS OF LC TESTING

NUMBER OF THESAURUS TERMS	5180
NUMBER OF MARC RECORDS	26000
NUMBER OF LC SUBJECT HEADINGS	30000

PROGRAMMING RESULTS

NUMBER OF WATER RESOURCE TERMS MATCHED TO LC SUBJECT HEADINGS	608 or 11.7%

are geography, oil industry, manpower, occupations, research and development and so forth.

3. No major sections of the LC stood out more significantly than any other.

4. There seems to be no guidelines for the use of plurals.

5. The LC could not be regarded as a metalanguage to be used for automated retrieval.

CHAPTER IX

SUMMARIES AND RECOMMENDATIONS

Increased demand for information and related services resulting from larger and more educated populations and more specialized technologies have resulted in new and unanticipated pressures being placed on today's information systems. The computer has been employed as a tool to help cope with these new pressures and to speed up existing procedures, many of which have not changed for a long time. However, even though the information handling procedures, being computerized, were not new, the machines could only be used after the information being handled by the system had been placed in a changed form, that is one suitable for computer manipulation. This reformatting has lead to the generation of numerous new tools which appear as large machine readable data bases.

Full advantage has not yet been taken of these new tools. In general they are being used as buffering devices making the information that is stored on them easy to update and allowing for that information to be printed in various formats. However, some information systems show more sophistication. For example, through the use of these new tools, the computer and machine readable data bases, detailed automatic searches of the information are now available.

This extension is only one of the possibilities offered by these new devices. Work should not stop here; in fact, the advantages that can be gained are limited only by the imagination of the information system designer [22].

One of the new uses of the tools is suggested in this thesis. It provides a method for measuring the suitability of existing classification schemes for a given subject area. Further, a judgment can be made about the degree of interrelationship of any thesaurus to an existing classification scheme when computer bases are available. If the classification is suitable and the relationship is satisfactory a combination of the thesaurus and the classification notation, resulting in a concordance, can be made. The majority of routine clerical tasks required when combining the thesaurus with the classification are completed through computer techniques allowing for more time to be spent on intellectual procedures. Through such computer techniques much assistance is offered to the subject specialist who will have to make decisions about the relationships of the subject descriptions.

Combining indexing with classification offers much flexibility and, if properly extended, would lead to a completely interactive system that can respond rapidly to changes in user needs and demands. For example, if word counts were taken, they could indicate when a new word should be entered in the indexing vocabulary. Continuous updating and use of the indexing language would indicate where changes are required in the classification.

During the late fifties and early sixties an automated system that handled large amounts of information and used manual indexing or classification as a means of identifying that information was considered to be very costly. These systems were expensive because classification was generally being performed by scientific subject specialists or analysts, often at the Masters or Ph.D. level. They were highly paid because it was difficult to attract them. At the time it was considered that no one else could do the work. This approach will doubtless be reconsidered today as other jobs for scientific specialists become more scarce. In addition, it is probable that the person with a general degree could be given the job of collecting, classifying and entering information into an automated system. In fact, studies such as those carried out by British researchers in the Medlar's system support this conclusion. Considerations such as those outlined above lead us to conclude that manual classification will continue even though it may be computer assisted.

The greatest difficulty faced by any system is that of insuring its correct usage. Users must be educated and taught how to effectively use their system, take advantage of it, and make criticisms. Although machine readable data bases and the computer offer great possibilities it is unrealistic to suppose that automated systems using these tools will come into widespread immediate use. The degree of use will be dependent not only upon the competent design of the system but also the efficiency of the related education program.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Heaps, D., "Problems and promises in information handling: the nature of information; its transfer and manipulation", Alberta Library Association Bulletin, vol. 2, no. 2, 19-30, March, 1969.

2. Klingbiel, P. H., The Future of Indexing and Retrieval Vocabularies, Defense Documentation Center, AD-716-200, November, 1970.

3. Heaps, D., Mercier, M., Cooke, G.A., "The Study of UDC and Other Indexing Languages through Computer Manipulation of Machine-Readable Data Bases", In: International Symposium: UDC in Relation to Other Indexing Languages, sponsored by Yugoslav Center for Technical and Scientific Documentation and FID, June 28-July 1, 1971, Hercg Novi, Yugoslavia, (in press).

4. Aitcheson, J., "The Thesaurofacet: A Multi-Purpose Retrieval Language Tool", Journal of Documentation, Vol. 26, No. 3, pp. 187-203, September, 1970.

5. Alber, F. M., On Line Thesaurus Design for an Integrated Information System, University of Alberta, Edmonton, Department of Computing Science, 1972. (unpublished master's thesis).

6. Heaps, D., and F. Alber, "Classifying, Indexing and Searching Resource Management Information via an On-Line Thesaurus", Western Canada Chapter, American Society for Information Science, Proceedings of the Annual Meeting, Oct., 1971. (Issued by Information Systems, University of Calgary, Oct., 1971).

7. Freeman, R. R. and Atherton, P., Final Report of the Research Project for the Evaluation of the UDC on the Indexing Language for a Mechanized Reference Retrieval System, Report AIP/UDC-9, May 1, 1968.

8. Information Systems Office, Library of Congress, MARC Manuals Used by the Library of Congress, Information Science and Automation Division, American Library Association, Chicago, 1962.

9. British Standards Institution, Guide to the Universal Decimal Classification, (UDC), British Standards House, London, 1963.

10. Mills, Jack, The Universal Decimal Classification, Rutgers Series on Systems for the Intellectual Organization of Information, vol. 1, New Brunswick, New Jersey, Graduate School of Library Service, the State University, Rutgers, 1964.

11. Interdepartmental Committee for Atmospheric Sciences Vocabulary, Washington, D.C., National Oceanic and Atmospheric Administration.

12. Heaps, D., et. al., Automation of a UDC Based Library for Searching Purposes, Paper presented at FID 2nd UDC Mechanization Seminar, Frankfurt, 1-5 June, 1970. (in press).

13. Ranganathan, S. R., Elements of Library Classification, Asia Publishing House, third edition, 1962.

14. Vickery, B. C., Faceted Classification, Rutgers University Press, New Brunswick, New Jersey, 1966

15. Freeman, R. R. and Atherton, P., File Organization and Search Strategy Using the Universal Decimal Classification in Mechanized Reference Retrieval Systems, Report AIP/UDC-5, September 15, 1967.

16. Heaps, H. S. and L. H. Thiel, "Optimum Procedures for Economic Information Retrieval", Information Storage and Retrieval, Vol. 6, No. 2, pp. 137-154, June, 1970.

17. Austin, C. J., MEDLARS, 1963-1967, Bethesda, Maryland, National Library of Medicine, 1968.

18. Foskett, A. C., The Subject Approach to Information, Clive Bingley, London, March, 1970.

19. Immroth, J. P., A Guide to Library of Congress Classification, Libraries Unlimited, Inc., Rochester, N.Y., 1968.

20. Schimmelpfeng, R. H. and C. D. Cook, ed., The Use of the Library of Congress Classification, American Library Association, Chicago, 1968.

21. Heaps, D., et al., "Search Programs for MARC Tapes at the University of Alberta", Western Canada Chapter, American Society for Information Science, Proceedings of the Annual Meeting, September, 1970. (Issued by Information Systems, University of Calgary, Dec. 1970).

22. Heaps, D., and W. Ingram, Computer Recognition and Graphical Reproduction of Patterns in Scientific and Technical Style, In: Proceedings of the 34th Annual Meeting of the American Society for Information Science, Denver, 1971. (in press).

B30015